STEPS TOWARDS A COMPUTATIONAL THEORY OF VISUAL MOTION DETECTION: DESIGNING A WORKING SYSTEM

by

Claude Lamontagne

Ph.D. Thesis

University of Edinburgh

1975

ABSTRACT

Creating a theory of visual motion detection involves considering particular motion detection systems and abstracting general information processing principles from them. But conversely, the design of any particular system should be set in as wide a theoretical framework as possible, to facilitate modifications, extensions, and comparisons with other particular systems. One part of the work reported in this dissertation is an effort to design a complete particular visual motion detection system, and to do it on the basis of computational concepts which cover the widest possible range of particular systems. The problem is tackled at two different levels: a higher level involving issues about when, where, and how motion detection abilities should be used in a complete visual system, and a lower level involving issues about what the motion detection abilities themselves precisely consist of and how they can actually be set to work. Solving the higher level issues yields the macro-structure of the particular motion detection system designed, and solving the lower level ones yields its micro-structure. The system's macro-structure is designed in the context of a monocular colour-blind system where bright line-drawings on dark backgrounds are the only stimuli allowed; discrete sampling of the light array falling on the retina (in both space and time) is also assumed. The system's micro-structure suffers the extra restriction of being designed in the context of two-dimensional motion only. The other part of the work reported in this dissertation consists in using the particular motion detection system designed and the general computational concepts used to do so as means of modelling biological visual motion detection systems. In this context, observations available from different fields of research are given new interpretations, available interpretations are strengthened and extended, and new evidence on the basis of new interpretations is provided, including a complete family of new visual phenomena providing an extensive experimental paradigm for investigating human eye-tracking behaviour and related phenomenal experiences. .

ACKNOWLEDGEMENTS

I wish to express my gratitude

-to my supervisor, Dr. J.A.M. Howe, for help and encouragement throughout the last four years,

-to Mark A., Raymond A., John C., Ben D., John K., Cyril L., David L., Roly L., Marilyn M., Hugh N., Tim R., Robert R., Jim S., Sylvia W., and other members of the departments of A.I. and Psychology for hints of all kinds and for creating a rich and pleasant research milieu,

-to Richard Young, for reading and commenting the first draft of this thesis,

-to Louise, for help and patience, and for sharing my faith in those aspects of this work which have not yet been "formally" validated,

-to the National Research Council of Canada, for financial support.

O investigator,

do not flatter yourself that you know

the things nature performs for herself,

but rejoice in knowing the purpose

of those things designed by your own mind.

Leonardo da Vinci
(transl. in L. Reti, 1974)

CONTENTS

INTRODUCTION

The roots of our investigation can be found in the very broad question: given a three-dimensional environment where physical objects emit, refract, or reflect light, and given an organism which has to adapt to this environment in some specified way, what would be the optimal way for this organism to compute or detect changes or movements in the environment through the medium of changes in the light array converging on any point of the environmental space?

Answering this question involves creating a computational theory of visual motion detection, where explicit comparisons can be made between all the possible systems in the chosen framework, and where the optimal one can be identified. The first step in creating such a theory is of course to find particular instances of motion detection systems, and to abstract general principles from them. Such particular instances can be obtained by unveiling the precise processes underlying biological visual motion detection systems and/or by designing artificial ones. These two different research attitudes can be found respectively in Physiology and Psychology on the one hand,

and in Computational Sciences (e.g. Artificial Intelligence) on the other. However, it appears that although very relevant partial evidence is available from these fields nowhere can we find a single complete particular instance of any non-trivial motion detection system (expressed in terms of explicit processes). The work reported in this dissertation is the fruit of four years of steady effort at trying to design a working visual motion detection system meant (on the long run) to detect a diversity of physical movements which is comparable to what the human visual motion detection system can handle, and to design this system in the largest possible context by considering at every stage of the task as many alternatives as possible, expressing both the problems and the solutions in terms of generalized computational concepts.

In our search for possible ways in which the different parts of our particular system could work we used to the fullest possible extent all the evidence that Physiology, Psychology, and Computational Sciences could offer. However, whenever a computational strategy was included in our system it was because this strategy was found computationally more suitable than any other one considered, not because it was presumably used by some particular biological system. This means that although we

are aiming at a motion detection system performing at a human level of sophistication we are not committed to the design of a system which necessarily uses the human visual processing strategies to reach this level of performance. However, if it is the case that we are not specifically attempting to simulate a human or other particular biological visual motion detection system, we believe that our system could and should be used at any stage of its development as a source of insights into how particular biological systems could work. Given any particular computational task it is indeed as ridiculous not to acknowledge the potential interest of artificial systems carrying out the task in generating hypotheses about how particular biological systems do carry it out as it is ridiculous not to acknowledge the interest of particular biological systems known to carry out the task in a search for possible ways of carrying it out.

In short, we can say that the work to be described in this dissertation bears on three main issues:

1-creating computational concepts which cover the widest possible range of particular visual motion detection systems;

2-using these concepts to articulate a complete particular working system;

3-looking for abilities of this particular system to

serve as computational model of visual motion detection in particular biological systems.

It was mentioned above that a "human level of performance" is aimed at in the design of our system. This cannot be considered as an extremely well defined goal since there is still a lot of discussion going on about what the human visual system is actually capable of detecting. However, we believe that enough is known about it, if only through introspection, to provide at least a rough indication of the kind of physical constraints which we ultimately wish our system to be able to detect. For one thing, "human level of performance" is well defined enough for us to say that the working visual motion detection system described in this dissertation does not reach such a level of performance, although we believe we are on the right path. Some important restrictions still have to be lifted before we can claim that our system has reached such a complex level of performance. The reason why we are emphasizing so much our concern for our system to reach (eventually) a human level of performance even though we have not succeeded yet in providing a working system performing at this level is that we believe that the ultimate level of complexity which any system can possibly reach is determined to a great extent by the very nature of the primitives, or building blocks, on which it rests, and

that if the ultimate aim is not well in mind when designing the most basic processing units of the system the whole enterprise might be completely jeopardized.

Now if we cannot claim to have designed a particular visual motion detection system performing at a human level of sophistication, how far have we got in this direction? Two answers should be given to this question since the design of our system was carried out at two different levels: a "higher" level where the more general issues about computing motion were discussed and where our system's overall structure (the macro-structure) was chosen, and a "lower" level where the problems of making the overall solution operational were discussed and where our system's underlying structure (the micro-structure) was chosen. The system's macro-structure is meant to cover such issues as when and where motion should be computed in the whole of visual processing, while the micro-structure is concerned with exactly how motion should be computed at the chosen stage of visual processing. Different sets of restrictions were imposed at these two different levels and they are as follows.

The system's macro-structure was designed in the context of a colour-blind monocular system where the eye contains a photo-sensitive surface, the retina, on to which a

single image is focussed at every moment. "All or none" responses to light intensities falling on the retina were assumed and legal stimuli were restricted to bright line-drawings on dark background. The practical implications of these restrictions can be grasped quite easily by realising that the restrictions are equivalent to those imposed on humans watching black and white movie films where objects are shown as bright line-drawings on dark background (as on a CRT screen for instance).

Given the above restrictions our system's chosen macro-structure should in principle be capable of allowing a human level of performance at detecting motion. However, no definite claim can be made until this macro-structure is made operational, and this requires the design of an explicit underlying structure, the micro-structure.

The system's <u>micro-structure</u> lies in the larger context of the macro-structure and therefore suffers the same restrictions. However, the task of implementing the macro-structure in its most powerful context was found to be too complex to be undertaken directly with any hope of success and an extra restriction was imposed on the design of the micro-structure. It was decided that a first reasonable step towards a human level of performance would

be to implement the macro-structure in the restricted context of two-dimensional motion. However, this would have to be done keeping in mind that the ultimate aim involves motion detection in a three-dimensional environment, which means that the two-dimensional motion detection system should be powerful enough to allow the eventual extension into the third dimension. There are a few rather controversial issues related to this possibility of reaching three-dimensional motion detection through two-dimensional motion detection, one of which has to do with the fact, which we fully acknowledge, that even though the retina only allows for a two-dimensional mapping of light stimuli there is absolutely no need for any actual system to go through two-dimensional motions in order to detect three-dimensional ones. Our belief that two-dimensional motion detection offers a valid and even a natural basis for three-dimensional motion detection is to be understood in the context of developmental or evolutionary requirements rather than in the context of some working visual system's hierarchical processing requirements.

Since it is rather hard for someone who is not familiar with the problems of motion detection to get a feeling for their complexity, especially in the restricted context of two-dimensional motions, it might be worth giving a few

examples. The main types of problems which the system described in the dissertation is designed to tackle are illustrated in the strip-cartoon shown in Figure 1. The successive frames presented in this strip-cartoon should be considered as representing an animated cartoon presented on a CRT screen (where Figure 1's black line-drawings on white background become bright line-drawings on dark background). One should imagine "continuous" events by filling in the gaps between the presented frames with a great many intermediate frames. Accelerations and decelerations are of course allowed and even welcome. Objects involved in translatory motions relative to the observer can be eye-tracked or not. At the start, the content of each frame should be thought of as a mosaic of independent "dots" of light or darkness.

Before going through a description of what our system is expected to detect when presented with the events portrayed in Figure 1 we want to be very clear about the fact that none of the physical objects represented in the cartoon are expected to be identified or recognised by our system as what they represent for the human observer. In other words our system does not "know" what is a lorry or a tree or a hot-air balloon, and when we say for instance that our system should "see" the lorry undergoing some global translatory movement we mean that our system should

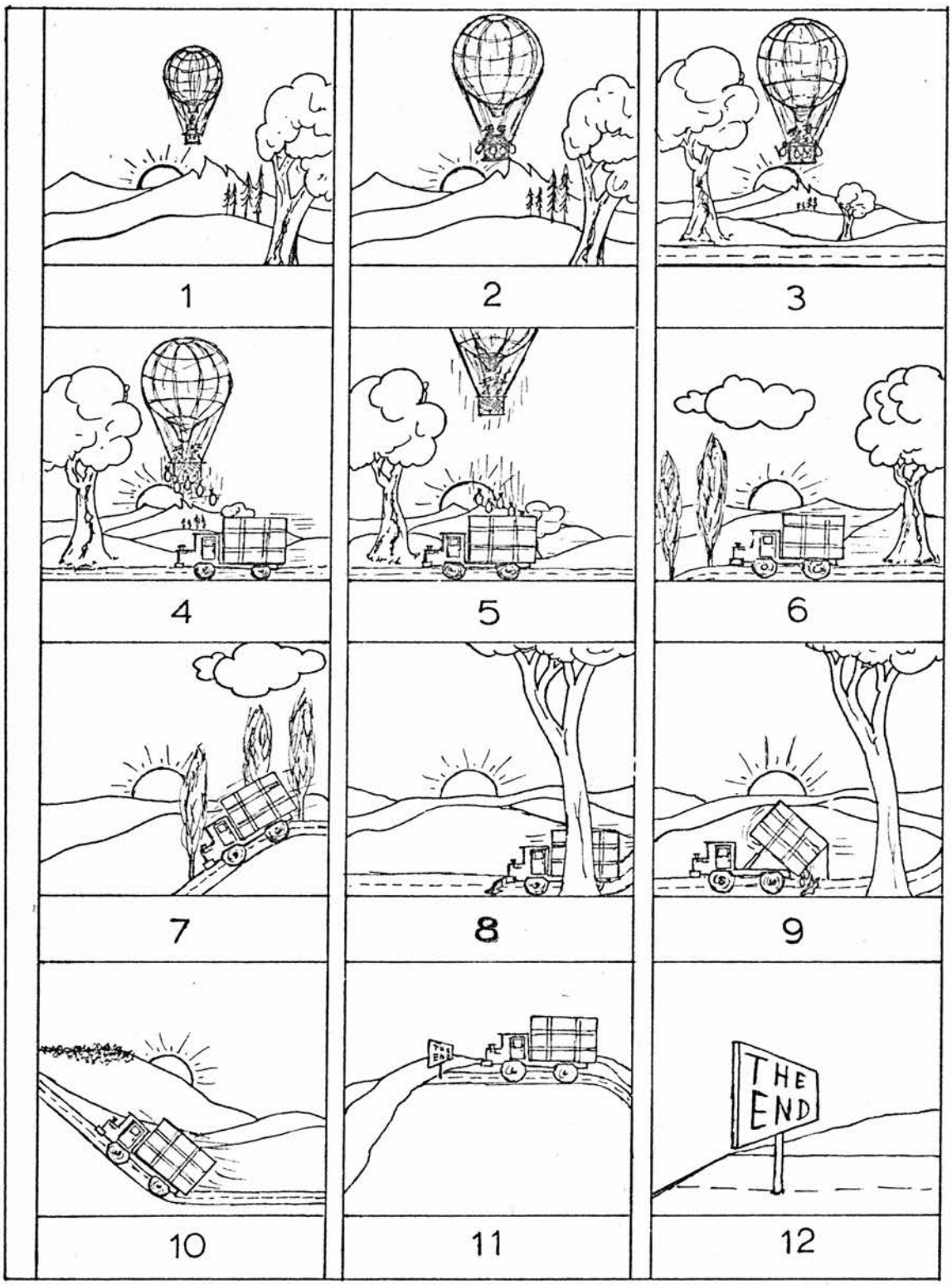


FIGURE 1. Some motion detection problems

"see" all the dots of light making up the lorry undergoing some global translatory movement. It is in fact on the basis of such information that higher level systems can actually put forward or strengthen the hypothesis that there is a lorry or a tree or a hot-air balloon out there, but this hypothesis or decision is not generated at the level of our system. Our system tackles motion.

Let us now describe the important features of the cartoon. From frame 1 to frame 2 all elements of the scene have stood still except for the balloon which has grown bigger relative to both the frame (i.e. the screen on which the animated cartoon is projected or the retina of the spectator) and the background (i.e. hills, trees, etc..). Our system should detect both facts. From frame 2 to frame 3 the balloon has carried on growing bigger, but relative to the background only; the balloon's size relative to the frame has remained exactly the same (1). Our system should again account for both facts. From frame 3 to frame 4 sand bags are dropped from the balloon

(1) Switching from an actual motion of the object (relative to the frame) to a motion of the background (relative to the frame) when showing an object moving about some environment is one of the most commonly used techniques in animated cartoons, and the transition is hardly noticed even by an attentive observer.

and move down as a whole (relative to both frame and background) while a lorry undergoing translatory movement (relative to both frame and background) enters the scene. Our system should be able to detect both global translatory movements (relative to both frame and background), not necessarily in parallel, and should be able to eye-track either one of the two moving "wholes". From frame 4 to frame 5 the balloon has moved up (relative to both frame and background) while the sand bags have moved further down falling into the lorry which has carried on its trip (relative to both frame and background). All three moving wholes should be identified as such by our system (one after the other), each being given a precise velocity relative to each of the two usual references, velocities relative to the frame (or system's retina) being possibly used to trigger eye-tracking.

From frame 5 to frame 6 the balloon has left the scene and the lorry (containing the sand bags) has carried on its trip relative to the background but has remained completely still relative to the frame. The sun has also "translated" relative to the background since it has remained still relative to the lorry; the two should then be seen as forming a <u>moving "whole"</u>. From frame 6 to frame 7 the lorry has once again undergone translatory movement relative to both the background (forward) and the

frame (backward), but on top of this it has also undergone a rotation relative to both references, and our system should detect this. From frame 7 to frame 8 the lorry has moved into a partly occluded position and has undergone similar types of translation and rotation as it did in the last step (having this time undergone a downward translation relative to the frame and a clockwise rotation). Our system should detect these motions of the lorry as a single whole even as the lorry gets partly occluded by passing behind the tree. From frame 8 to frame 9 the lorry has once more undergone translation relative to both references, but the rear part of the lorry has also started to translate and rotate relative to the lorry itself, as well as relative to the lorry's background and the frame, unloading the sand bags into a pot-hole in the road. Our system should be able to detect movement at any one of these levels. From frame 9 to frame 10 the lorry has once again translated and rotated relative to both references but the interesting point is that the part of the lorry which has moved throughout the last step has remained completely still relative to the frame while resuming its original orientation and position relative to the lorry. Our system should be able to notice this. From frame 10 to frame 11 the lorry has carried on translating and rotating as a whole relative to both references and has come to a stop in front of a road

sign. As soon as the movement stops the lorry should cease being considered as an independent whole by our system and the whole scene should become a single unit standing still. Finally, from frame 11 to frame 12, the whole scene grows bigger as a single unit, and should be detected as such by our system.

The dissertation will consist of two parts. Part I will be devoted to the system's macro-structure. It will contain three chapters. In the first (Chapter I) we will set the context for the discussion by presenting basic assumptions and terminology concerning visual information processing in general, and by proposing one precise way of looking at the motion detection problem and its possible solutions. In the second (Chapter II) the respective contributions of Artificial Intelligence, Physiology, and Psychology in the context of our problem will be assessed. In the third (Chapter III) our own solution, i.e. the system's chosen macro-structure, will be presented and some of its implications will be discussed.

Part II will be devoted to the design of the system's micro-structure, or set of precise mechanisms implementing the general solution chosen in Part I. It will consist of three chapters. In the first (Chapter IV) some general-purpose primitive concepts will be created to

serve as building blocks in the design of the micro-structure. In the second (Chapter V) the actual micro-structure will be described. In both chapters problems and solutions will be presented at an intuitive level before being embodied into precise nerve nets. In the third (Chapter VI) the system's suitability for modelling aspects of particular biological visual motion detection systems will be discussed.

Our hope is that our verbal account of a visual system designed mostly on the basis of visual concepts will be sufficiently precise to allow the reader to climb back up to the visual level and experience the visual life of the system: we found the experience well worth all the work which was put into making it possible, and even much more.

PART I

THE SYSTEM'S MACRO-STRUCTURE

or

Where the heart of the problem

is exposed

and one solution proposed

CHAPTER I

The problem

I.1 Preliminaries

Stating the problem of visual motion detection requires the context of an overall visual system. In this first section we will be concerned firstly with presenting some basic assumptions concerning the visual system within which our attempt to state and solve the motion detection problem will be set, and secondly with introducing basic concepts (and associated terminology) covering the different possible aspects of information processing within such a system with an emphasis on those aspects more directly concerned with the motion detection issue.

We have already said that the visual system would be assumed to be monocular and to react in an "all or none" manner to light intensities projected on to its retina from some stimulus structure consisting of bright line-drawings on dark background.

Our first problem is to decide how the system should sample the light array projected on to its retina.

Our system will be assumed to work on the basis of discrete occurrences of light, i.e. the detection of the "flow of light" coming to the system's eye will first involve breaking this flow spacewise and timewise into a number of discrete units.

In the spatial domain this can be done by using a retina consisting of a two-dimensional array of light-sensitive cells (or receptors) where a signal fired by anyone of these cells specifies a particular position of light occurrence. It will be sufficient for a start to assume that the retinal receptors are tightly packed together without overlapping each other (thereby implementing what we call an "adjacent" type of sampling).

In the temporal domain discreteness can be achieved by checking for the presence or absence of the "position" signals within a specified finite period of time, the belonging of any position signal to any one of these periods specifying a particular moment of light occurrence. We will assume that our system's temporal sampling strategy involves sampling for a short period of time after which a no-sampling period, or sampling "gap", is allowed to go by before the next sampling starts. This type of sampling (which we call a "gap" sampling) offers great conceptual simplicity (it is in fact the equivalent

of a cine-camera type of temporal sampling) and although it involves losses of potential information (because of the sampling gap) it remains sufficiently powerful for our purpose. A discussion of the different possible types of discrete sampling strategies (in both time and space) and of their respective merits and weaknesses can be found in Appendix A.

Now the visual "units" created through these sampling strategies and unequivocally characterized through a position on the retina and a moment in the processing can be considered as proper "visual entities". They are in fact the most primitive visual entities in the system and they will be the ultimate basis of all processing taking place in it. They will from now on be referred to as atomic visual entities (a.v.e.'s).

A.v.e.'s can be thought of as representing "points" or "dots" of light falling on the retina at particular moments (a dot being exhaustively described through a position and a moment), and since line-drawings on dark background are the only legal stimuli for our system a.v.e.'s can be thought of as representing directly contrasting "points" or "dots" in the visual scene.

Assuming the restrictions introduced so far we view the

task of any visual system as being one of

1- grouping visual entities (starting with a.v.e.'s) under well defined criteria into higher level visual entities,

2- characterizing the visual entities so obtained with well defined features bearing on each one of these entities as a whole,

3- repeating this process until visual entities are obtained which represent and qualify adequately the physical environment.

Since the analysis has to start with a.v.e.'s let us have a look at them first and see what the task implies at their level. First, even though a.v.e.'s are our starting point they were in a sense obtained by using "grouping" criteria, namely "existence of light" in a particular "portion of space" within a particular "portion of time", and these criteria were also used as characterizing features for a.v.e.'s (each a.v.e. being described in terms of its position and moment). This shows that the minimal characterization of a visual entity can be found in the grouping criteria which were used to generate it. However, characterization can usually reach much beyond the features used as grouping criteria, and it is on the basis of the characterizing features of the visual entities lying at some level, whether or not these

features include those which were used as grouping criteria to reach this level, that the grouping criteria to reach yet a higher level of visual entities will be defined. So let us now take a closer look at the type of features which our system is likely to use for characterization purposes.

As we have said above, grouping criteria can also be used as characterizing features; we will call these features "critical characterizing features" in contrast with the rest of the characterizing features which will be called "incidental characterizing features".

Now any characterizing feature, whether it is critical or incidental, is either single-valued or multi-valued. A single-valued feature can be thought of as being a predicate, i.e. an attribute or feature which is either true or false given some particular visual entity. On the other hand a multi-valued feature is one which takes one value or another from some domain, depending on the visual entity (hereafter v.e.) being characterized. For instance "colour", "intensity", "moment", "position" are multi-valued features, while features like "having two adjacent neighbours", "convexity", "triangularity", "45 degree orientation" are single-valued features. The most interesting of the two types of feature for us is the

multi-valued one because it presupposes, for each such feature, a "domain" or "dimension" within which comparisons can be made and new relations created. If a system reaches a stage where there is but a single v.e. characterized with but single-valued features, the grouping is over. Such a case could be trivially approximated by having but a single a.v.e. to play with. "Occurrence of light" is obviously an "all or none" type of feature, and so are "in a particular position" and "at a particular moment"; notice here that the last two features mentioned are "singularizations" of two multi-valued features, namely "position" and "moment", since each a.v.e. is defined as being specific to one given position and one given moment. So with a single a.v.e. no system can do much. However, with many a.v.e.'s in space and in time these two "multi-valued" dimensions, made available to the system by its basic sampling strategies, provide the system with two main ways of defining grouping criteria and characterizing features, whether or not they get to be single- or multi-valued, and this brings up the last, but most important, distinction which will be made in the context of characterizing features, the distinction between frozen and running features.

Any characterizing feature which rests on values of

features characterizing CONTEMPORARY lower level v.e.'s will be called a "frozen feature", and any characterizing feature which rests on values of features characterizing NON-CONTEMPORARY lower level v.e.'s will be called a "running feature". Frozen features therefore characterize groupings through space while running features characterize groupings through time.

Let us consider the case where a "continuous" straight line (parallel to the plane of the retina) is projected on to the retina at moment-0 and covers nine receptor-cells. Within this moment-0 there are nine receptors firing signals, specifying nine different positions of light occurrence all in moment-0, and giving birth to nine different a.v.e.'s. Now since a.v.e.'s are (by definition) the result of a grouping taking place within moment-0 alone, they can only be characterized by frozen features; so "occurrence (or existence) of light", "position", and "moment", are three instances of frozen features. Any higher level feature derived from some or all of the values of the features bearing on these a.v.e.'s alone will itself be a frozen feature. For instance, finding out that the detected positions of light on the retina are adjacent, in a straight line, in some orientation, and that there is a certain number of them creates as many new frozen features, all of which could be

used to characterize the more global v.e. consisting of all nine local a.v.e.'s. Notice here that two of these new features (adjacency and straightness) are single-valued, while the other two (orientation and size) are multi-valued; this means that the v.e. which they characterize is "invariant" under changes of size and orientation. We would have to know what the critical characterizing features of this v.e. are (i.e. what the grouping criteria were) in order to decide if "straightness" and "adjacency" are essential to its existence or not, but this does not matter for the moment. The main point here is that all of these features rest on values of features characterizing lower level v.e.'s (namely a.v.e.'s) which all belong to the same moment (namely moment-0), and this means that they are all frozen features.

Now instead of the nine a.v.e.'s being detected within but a single moment (moment-0) let us consider the case where the same positions of light occurence are detected, but the light is made to "occur" in one position only within each moment so that all nine positions are <u>successively</u> hit by the light, one position per moment. This means that overall we will still have nine a.v.e.'s created, but this time their <u>moment</u> of existence as well as their position of existence will be the differentiating factor

between them. In this situation we can introduce a completely different type of grouping with its particular type of features. In our discussion above we created "frozen groups" and discussed the associated frozen features. Here we will discuss "running groups" and the associated "running features". A running group, like any group, is obtained by creating a global v.e. out of more local ones, but in this case "global" and "local" are to be understood in the temporal dimension. If we consider moments 0 and 1 of our example we have two a.v.e's (adjacent in time as well as in space) which can be grouped through time into one single v.e., possibly under the criterion of "temporal adjacency", and possibly characterized with such running features as a direction and a speed of movement. (An interesting point here is that if "temporal adjacency of a.v.e.'s" is chosen as grouping criterion, then "particular position" will not be for the new v.e. the <u>critical</u> feature which it was for an a.v.e., and this opens up explicitly within the scope of this new v.e. the whole range of values offered by position as a multi-valued feature.) The point here is that a v.e. with velocity is temporally more global than an a.v.e. or any other v.e. limited to a single moment, however complex its frozen structure might be. Similarly a v.e. characterized with an acceleration and deceleration is temporally more global than one which is

only characterized with a primary speed, because it covers periods of three moments while the latter covers periods of only two moments. So as our nine a.v.e.'s succeed each other in time they can be grouped into higher and higher level objects in a way analogous to what can be done with them in an essentially spatial, or frozen, context. The examples given here might tend to suggest that "translatory motion" is really what running features are all about. We wish to stress that although it is of course the case that translatory motion is a running feature, not only do running features include other types of motion but they also include other features which can by no means be called "motion" features. Running features can indeed be derived from values of any other features, either single or multi-valued, while actual motion can only be computed from certain multi-valued features, the actual feature determining the type of motion, and its pool of values providing the actual velocity detection space.

Now an interesting point is that although a running feature can be derived directly from values of another (more local) running one all running features can be traced back to some point of their computational history where values of some frozen feature are taken into account. The furthest such point for any feature is of

course the one moment where the analysis starts, where everything is frozen, where a.v.e.'s are given birth. For instance a translatory acceleration is based on translatory speed, which is another running feature, but this one is based on "position" which is a frozen one. This fact seems to be the intuitive ground on which people stand when saying : "motion alone cannot exist; there has to be something moving". But the validity of this intuitive notion is often (not to say always) jeopardised by explicit comments which make one think that the "something" which is referred to has to be the visual equivalent of physical objects in a scene. And this is where powerful objections can be thrown in : there is no need for a v.e. to have attained the level of correspondence with physical objects, or anywhere close to them, in order to claim or allow running characterization including motion features - in fact, two a.v.e.'s in different moments and positions are sufficient to get motion computed. This weakens considerably any standpoint assuming that frozen groupings have to be exhausted before running groupings can be considered. There is no computational necessity to assume such a total dependence of running groupings on frozen ones. In fact the "right to existence" of running features as a unique and relatively independent type of characterization can be fought for in two different contexts: firstly, in the

context of a working visual system, like the one we have just started to discuss, where frozen and running features interact in the best possible way to provide the system with an adequate interpretation of its physical environment; and secondly, in a developmental or evolutionary context where frozen and running features interact in the best possible way to create new levels of analysis, with new and more global features, to provide the system with an adequate way of interpreting its physical environment. However, since our main purpose in this dissertation concerns the role of running features in the first context only we will concentrate on this one for the time being. A discussion of the role of running features in the second context, the developmental one, can be found in Appendix B.

In the usual context of a working visual system, running features can be made to play either one of the two roles which characterizing features can play : (1) they can be used simply as characterizing features and/or (2) they can be used as grouping criteria. The first type of use is totally independent of the level of analysis which has been attained : running features can be made to characterize v.e.'s as soon as local v.e.'s exist in two different moments. The important point here is that running features, and thereby motion features, can

potentially bear on any v.e.'s detected through time, which is a much wider set of v.e.'s than the set of those which correspond to physical objects in any visual scene. This means that running features, including motion features, can start being taken into consideration long before the system has got any idea of which physical entities are present in its field of view; in fact we feel that as soon as one considers grouping two a.v.e.'s which have the same moment but different positions one should also consider grouping a.v.e.'s which are distinct by their moment rather than or as well as by their positions. So although it remains true that there has to be a frozen feature at the basis of any running one it nevertheless is the case that there is no reason why the system should wait until the complete frozen analysis of a scene is achieved before worrying about its running analysis; on the contrary the system has every reason, as will be seen later, to start the running analysis as soon as it can, and this is very early in the process indeed, just next to a.v.e.'s. The second possible use of running features (i.e. as grouping criteria) goes even further in liberating running analysis from frozen analysis by stating that not only can running features (including motion features) exist long before the frozen analysis is completed but they can also play an active role in achieving this frozen analysis itself, thereby creating

situations where frozen features depend on running ones. These situations are to be found in cases of frozen grouping on running grounds. Such cases occur for instance when the system considers a set of v.e.'s which move together and groups them under the criterion of "common velocity", characterizing the group with frozen features such as "closure", or "connectedness", or "shape" (e.g. a "V-shaped" group of flying birds). Running features can of course also be used as criteria for running groupings (i.e. running groupings on running grounds), and this possibility, although it does not have for our argument the advantage of enslaving frozen features, at least stresses the relative independence of running features from frozen ones.

In short, we assumed discrete sampling of the light array falling on our system's retina, opting for an "adjacent" type of spatial sampling and a "gap" type of temporal sampling. This decision to split the detection of light into units of space and units of time provided the basis for identifying the basic material on which visual processes are to be set to work as being visual entities whose existence is essentially characterized by a precise position in space and a precise moment in time; they were called "atomic visual entities" (a.v.e.'s). The task of the visual system was then said to be concerned with

grouping a.v.e.'s into higher level visual entities (v.e.'s), a visual entity being any particular set of more local visual entities given some characterization as a whole, using as grouping criteria and global characterizing features the local characterizing features (and/or their derivatives) of the visual entities being grouped; this grouping process should carry on until an adequate description of the physical scene under analysis is reached. The characterizing features of any v.e. were mainly classified as being either running or frozen, and either single -valued or multi-valued. This is where motion was allowed in, as a particular kind of running feature resting on multi-valued features whose values hold some detectable relations between them.

I.2 Visual motion detection

Now that the visual information processing context in which motion detection processes should be set has been defined we can concentrate on motion detection itself.

In the forthcoming discussion we will be mostly concerned with "running groupings". These "groupings through time" are the very basis of motion detection. However, to make the discussion more appealing intuitively we will talk about motion detection as "motion detection" rather than as a more formal "running grouping". Actually we will restrict the use of the term "grouping" (and others of its family) to exclusively <u>frozen</u> groupings, so that the v.e.'s referred to will all be groups of more local v.e.'s <u>belonging to the same moments</u>, that is to say, frozen groups' succession through time will be our main concern. For instance, when in due course we talk about the "group first and compute motion afterwards" solution, we will strictly be speaking of a "do <u>frozen</u> grouping first and <u>running</u> grouping afterwards" solution. Once motion detection has been thoroughly understood, its actual role as a running grouping can be best appreciated and the generality of the concept of "visual grouping" can be endorsed. Until then we will ignore the fact that

"groupings" can be done in time in favour of adopting "motion detection" as being what happens in time.

In motion detection the most acute problems seem to lie in the IDENTIFICATION of corresponding v.e.'s as they appear through successive moments, and in the SPECIFICATION of the transformations which they undergo from moment to moment. That is to say, given two sets of a.v.e.'s (set-0 and set-1), set-0 having been detected at moment-0 and set-1 at moment-1, how could a visual system (a) match adequately a.v.e.'s in set-1 with a.v.e.'s in set-0 and (b) represent adequately the possible transformations in the characterizations of a.v.e.'s from set-0 to set-1 ?

Since "changes through time" are what transformations are all about, and since multi-valued features are features which can bear changes in value from moment to moment, the specification of transformations of a v.e. (question (b) above) can best be achieved through the use of multi-valued features characterizing this v.e.; we will therefore call the process of detecting values of appropriate multi-valued features specifying the transformations of the v.e. the M-CHARACTERIZATION process (M for Motion or Multi-valued). By contrast the identification problem (question (a) above) can best be solved by using single-valued features to specify each

v.e. (these features being "transformation-free"); we will therefore call this process of detecting values of appropriate single-valued features specifying the identity of a v.e. the S-CHARACTERIZATION process (S for Still or Single-valued). The whole issue of motion detection is of course to determine exactly how our visual system should go about S-characterization and M-characterization.

To establish ideas on these important issues let us consider some concrete cases. These will be scenes of events involving two moments only. Unless specified otherwise each scene will be presented graphically within a single frame representing a "retinal" grid of receptors on which each a.v.e. detected at moment-0 will be marked with a small circle and each one detected at moment-1 with a small cross. The desired outcome of the analysis of the two sets of a.v.e.'s by the visual system will be stated under each scene in terms of vector arrows and verbal comments.

In Scene I (Figure 2) both S-characterization and M-characterization are trivial since there is only a single a.v.e. (in each moment). This v.e. is readily identified at all times (being S-characterized through its very "existence") and is easily M-characterized through the most obvious multi-valued feature at this level,

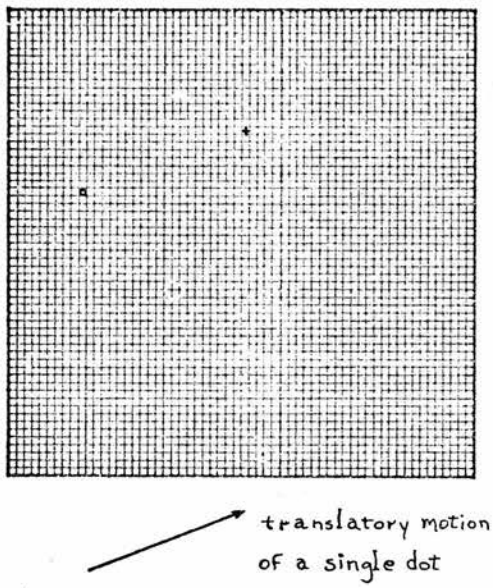


FIGURE 2. Scene I.

namely position (1). However, while M-characterization is achieved through the detection of position, the precise specification of the change in position, or translatory movement, occurring in Scene I has yet to be made. This "precise specification" is of course the velocity of the movement, which should be specified in terms of (a) a direction of movement and (b) a speed of movement. However, the processes underlying velocity computing will not be discussed until later, the priority being given to the more general problems of S- and M-characterization.

(1) Notice here that the more formal meaning of this is that for the purpose of M-characterization "position" has become multi-valued : we are indeed not talking of a single a.v.e. here, but of two a.v.e.'s grouped in time under the criterion of "existence" only.

To sum up our concern for Scene I we can say that it was mainly directed at an acknowledgement of the problem of velocity computing as a genuine but secondary aspect of the present discussion. The simplicity of the scene could not allow for an interesting discussion of much else, the desired interpretation of the scene (i.e. "a single translating dot") being obtained without much effort because of the existence of a single object whose simplicity as an a.v.e. did not allow for much ambiguity.

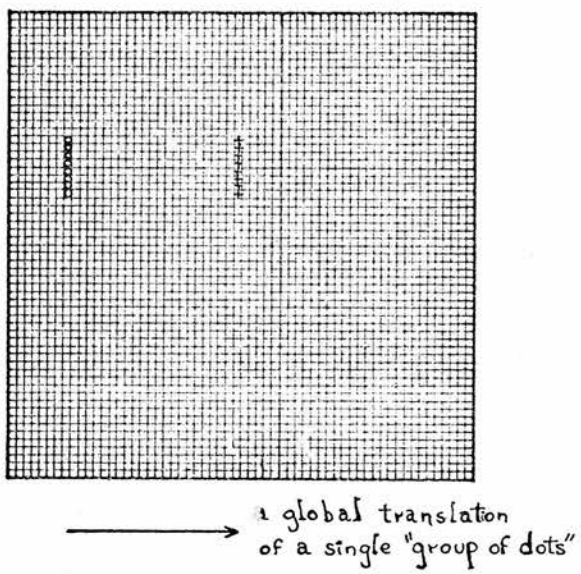


FIGURE 3. SCENE II.

With Scene II (Figure 3) the presence of many a.v.e.'s (in each moment) raises more interesting issues. The question arises as to how the a.v.e.'s should be grouped (since the desired interpretation mentions "a single group of dots") and on which grounds this should be done, as well as the

question of deciding if S-characterization and M-characterization (i.e. motion detection) should be carried out before or after the grouping, and how they should be carried out.

Before going any further let us recall quickly how (frozen) "grouping" and "detecting motion" (i.e. S- and M-characterizing) differ from each other and how they are made to relate in the overall process. As seen in Section I.1, "grouping" and "global characterization" are the general processes by which the visual system achieves a more global representation of the set of v.e.'s which it has succeeded in representing at any given level of analysis (the "absolute" starting point being the level of a.v.e.'s). (Frozen) grouping processes and motion detection interfere little with each other. (Frozen) grouping processes concern contemporary v.e.'s only and depend on motion only to the extent to which they need it as criterion to group some set of (contemporary) v.e.'s. Now as far as motion detection is concerned the main point regarding its dependence on (frozen) grouping processes is that motion has to be computed <u>at some level of (frozen) grouping</u> (i.e. after so many groupings have already been achieved, and before so many others are carried out), S- and M-characterization necessarily having to be carried out <u>on the v.e.'s available at this level</u>. Therefore, an

important problem to be solved is at which level(s) of grouping should the visual system compute (such and such a) motion?

For the precise case of Scene II (Figure 3) the question becomes : (a) should the visual system group first all seven a.v.e.'s in each moment (using as criterion for instance the frozen feature "adjacency to another a.v.e.", derived from primitive local positions, i.e. any a.v.e. which is adjacent to another a.v.e. will be part of the new v.e.) and then compute motion by trivially identifying (as was the case for Scene I) the only existing v.e. at any moment, M-characterizing it (also at every moment) with a single global position whose change from moment to moment, once processed by the velocity detection structures, will complete the desired description of a "translating group of dots moving towards the right at speed S; or (b) should the visual system instead compute motion first, at the level of a.v.e.'s, and then use the computed velocities (which will all be the same if Identification is well done) as criterion for grouping the seven moving v.e.'s into a single more global v.e. moving with a global velocity simply derived by giving a global status to the local velocity of one of the local v.e.'s grouped, thereby achieving the desired representation of the Scene? With this second strategy, however,

Identification is non-trivial since many entities, seven altogether, have to be considered individually. A possible Identification strategy, the "proximity" strategy, is to match each v.e. detected at moment-0 with the v.e. detected at moment-1 which stands closest to where the moment-0 v.e. stood, but it is important to realise here that such an Identification strategy is rather weak, being based on the very loose S-characterization of a single-valued feature (i.e. proximity) which is derived directly from the potentially multi-valued feature "position", whose very nature is highly inappropriate in the context of strict object identity. This is however the only alternative we have in the present case since no other S-characterization is available which could offer sufficient specificity to each individual v.e. to allow Identification on a stronger basis (all v.e.'s being "dots"). Whatever the case may be the "proximity" strategy is perfectly suitable for Identification purposes in Scene II, and since furthermore M-characterization is in this case also quite adequately and even trivially carried out as for Scene I, the "compute motion first" strategy is quite valid for Scene II. With Scene II we therefore have two adequate and quite different types of solutions : the "group first" type and the "compute motion first" type. The simplicity of Scene II does not however allow the critical evaluation

of these solutions' respective weaknesses and virtues, the main purpose of this Scene having been simply to introduce the two main ways of going about solving the problem of motion perception. Let us now discuss the relative merits of the two types of solution.

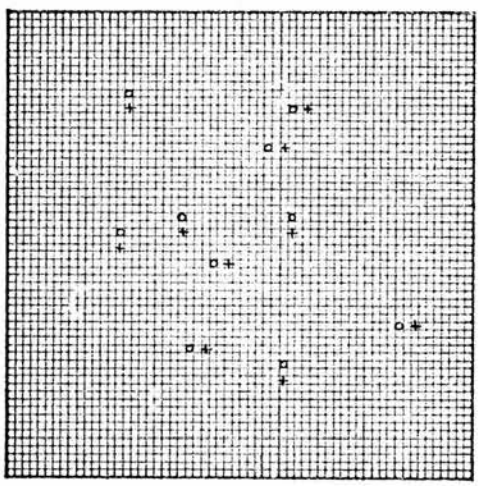


FIGURE 4. Scene III.

Scene III (Figure 4) is interesting in that the "motion first" solution suits the situation beautifully while the "group first" solution is found to be almost totally inadequate. Indeed, a grouping on purely frozen grounds (before getting down to compute any motion) would need to stick so tightly to local information (in order to allow for proper M-characterization) that the enterprise loses all its appeal. To appreciate this problem one has to try assigning multi-valued global features to the set(s) of

a.v.e.'s at any moment in such a way that the system manages, by considering the successive values of these global features only, to reach the desired interpretation of "two distinct groups of dots, one moving right at speed S1 and the other one moving down at speed S2". On the other hand, if motion is computed directly on a.v.e.'s (identified through the simple "proximity" strategy already mentioned, and M-characterized through the most obvious multi-valued feature at their level, i.e. retinal position) the desired interpretation is easily obtained by grouping together all those a.v.e.'s which have a common velocity, this velocity then being given a global status as regards the group of a.v.e.'s which it specifies.

A very important point about "Scene III types of situations" is that far from being queer cases thrown in only to muddle the issue they represent the very basis of the problem of "body identification" in scene analysis in as far as running features are concerned. These situations can be grouped under the label "movement field effects" (or "running field effects") and can be found frequently in every day life, from the occasional swaying of tree leaves in the wind to the ever-present motion parallax of just about every object in our three-dimensional physical environment.

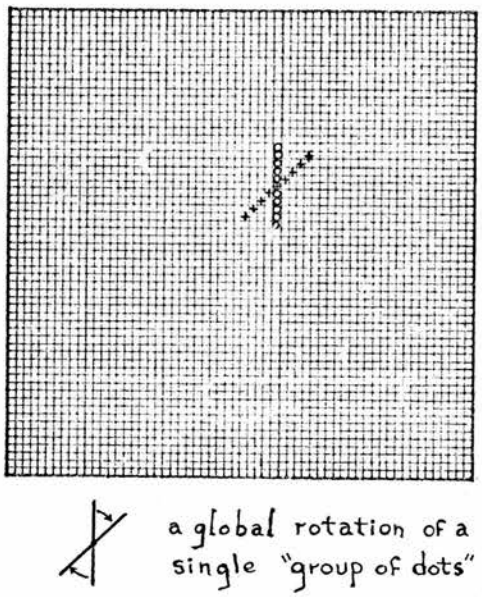
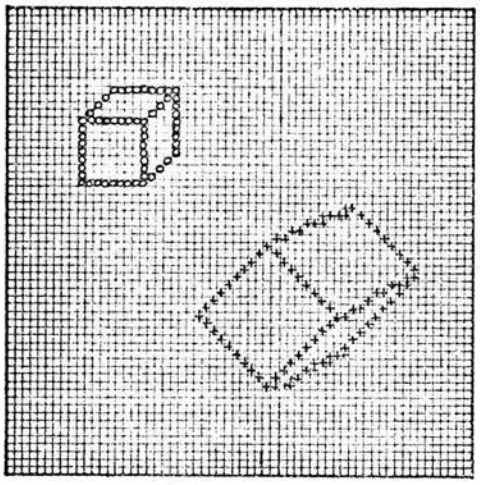


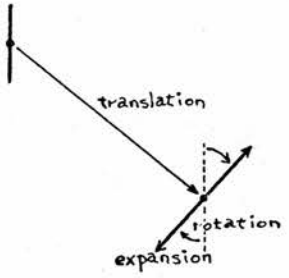
FIGURE 5. Scene IV.

Now, on the basis of the above discussion about Scene III, if one wants to adopt the "compute motion first" solution as overall strategy, Scene IV (Figure 5) introduces a difficulty. Indeed, in Scene IV, the "compute motion first" solution implies a rather complex grouping of the moving a.v.e.'s in order to achieve the desired interpretation ("a rotating group of elements"). The problem comes from the fact that if one wants to use the movements of a.v.e.'s to define the criterion under which to group them, the "common atomic translatory velocity" criterion (which worked well for Scenes II and III) obviously does not work if the group of a.v.e.'s is not itself undergoing a global translatory movement (like in the case of Scene IV where it is undergoing a rotatory movement). In the case of Scene IV, if we want to use

atomic velocities as basis for grouping, the grouping criterion has to embody mathematical descriptions of relational velocities of elements belonging to a whole undergoing rotation, requiring such things as tangential movements and other "complex" features. One does not have to go very far into the solution just hinted at before realising that although it sounds feasible it is dramatically more complex and less natural than the "group first" solution which would simply group the eleven elements detected at any moment of the event portrayed in Scene IV into a single group under, for example, the frozen criterion of adjacency. This new global visual object would then be M-characterized through the multi-valued feature "global orientation" (worked out under the frozen criterion of "a certain type of adjacency") which, computed at every moment, would yield the changing values required to compute the rotatory velocity of the trivially identified single visual object (S-characterized through the trivial feature "existence"), thereby achieving the desired interpretation.

Scene V (Figure 6) goes even further than Scene IV by showing that in some cases it is virtually impossible to apply the "compute motion first" solution. In the case of Scene IV this solution only became rather clumsy at the final stage of the process, i.e. when the time came to





a single group of dots
translating, rotating
and expanding

FIGURE 6. SCENE V.

group the a.v.e.'s on the basis of their velocities into a more global entity with a global velocity, the Identification and M- characterization problems having been solved quickly and easily. With Scene V difficulties arise even with Identification, making it virtually impossible at the level of a.v.e.'s. Indeed there seems to be no way in which one could go about <u>locally identifying</u> each a.v.e. detected at moment-1 as corresponding to such or such an a.v.e. detected at moment-0. In a sense we can say that Scene V is the converse of Scene III (figure 4), the latter having shown the complete inadequacy of the "group first" solution in at least one case while the former is showing the complete inadequacy of the "compute motion first" solution in also at least one case. The inadequacy of this solution is

however well compensated for by the ease with which the "group first" solution handles the scene, although M-characterization gets to be a little more difficult than for Scene IV, requiring three different global (multi-valued) features to be computed (namely global position, global orientation, and global size), while Identification becomes trivial by the fact that the grouping (carried out for instance using the "adjacency" criterion already talked about) comes up at any moment with a single global entity (containing all a.v.e.'s). This Identification strategy together with the three M-characterizing global features easily provide the basis for obtaining the desired interpretation of "a group of dots translating with a velocity T, rotating with a velocity R, and expanding with a velocity E".

What makes Identification at the level of a.v.e.'s so hopeless in the case of Scene V is of course the large spatial step imposed on a.v.e.'s from moment-0 to moment-1. In fact there is not even need for such a large "gap" in order to fool an Identification strategy based on spatial proximity through time, a critical gap only having to be sufficiently wide to place a few moment-1 a.v.e.'s further away from the moment-0 corresponding ones than from any other ones. But, of course, however small the gap, such situations involve non- continuous events, and

for this reason could be argued to be a non-legitimate dimension of the problem, some people feeling that such movements only exist in artificial situations which need not concern those interested in "reality". That such non-continuous events are of great importance to the definition of our problem should become obvious from the two following arguments. The first argument bears on natural stimulus structures allowing for non-continuous events which a visual system should be designed to analyse in terms of proper continuous movement. A first case involves an object moving in an environment rich in occluding objects, like a fox running in a forest, a person moving through a crowd, or a boy running behind a picket fence, while a second case involves an object at rest which suddenly and quickly moves to another state of rest, like an eye winking, or a head turning suddenly. This second case is in fact only meaningful relative to the temporal sampling of the perceiving system (the speeds at which the eye-lids have to close and the head has to turn in order to make the argument valid being highly dependent on the temporal sampling period of the visual system concerned), which brings us to our second and main argument. This argument bears on the temporal sampling strategy(ies) used by the visual system concerned and stresses that a case such as Scene V could be the result of a "cine-camera" type of sampling (or "gap" sampling) as

much as the result of the stimulus structure itself. It is therefore argued that if one is not to limit the possible solutions of the problem to "continuous sampling" (which anyway, as argued in Appendix A, is quite a slippery position to adopt), situations such as Scene V have to be taken into account.

One of the main outcomes of our discussion so far is the realisation that there are at least two rather different ways of achieving the representation of a globally moving figure (or v.e.), namely by grouping on the basis of atomic motions or by grouping on purely frozen grounds before computing global motion as such, and that while one or the other was always found to be adequate whatever the situation considered, neither of them could be found to be adequate (or even applicable) in all situations. Another important outcome of the discussion is the realisation that Identification and M-characterization, i.e. motion detection as such, are more or less of a problem depending on the level of grouping at which they are applied (this level of grouping depending on the grouping strategy), and that furthermore the respective degrees of complexity of Identification and M-characterization at any given level of grouping are inversely proportional to each other. This last remark is not very surprising if one realises that for instance at the lowest level of grouping (i.e.

the level of a.v.e.'s) v.e.'s to be identified are many while their potential features for M-characterization are few (just about only a position in fact), while at the highest level of grouping (i.e. the level of a single global v.e. "containing" all a.v.e.'s) v.e.'s to be identified are few (there is only one in fact) while their potential features for M-characterization are many, and have to be many to account for all possible movements. This boils down to realising that since M-characterization is by definition concerned with the specification of features which cover the range of possible changes within a given v.e., the more elaborate this one is (i.e. the further away it stands from the level of a.v.e.'s) the heavier M-characterization has to become; conversely, since Identification is by definition concerned with keeping track of v.e.'s from moment to moment, the more numerous these are (i.e. the closer they are to the level of a.v.e.'s) the more tricky Identification becomes. The problem of Identification is, however, a little more complex than that since what really makes it hard (and most of the time even impossible) to apply at the level of a.v.e.'s is the absence of powerful single-valued features for S-characterization. We have in fact said very little about S-characterization since we have started discussing scenes. The reason for this relative silence lies in the fact that the main levels of

grouping which were adopted were either the a.v.e.'s level, where S-characterization had to be done on the grounds of "proximity", or the level of a single object (or v.e.) consisting of some grouping of all a.v.e.'s, where S-characterization was achieved on the basis of a trivial but perfectly adequate single-valued feature, namely the "existence" of the single v.e. at any moment. This last attitude of grouping a.v.e.'s right up to the level of a single global v.e. before computing motion sounded intuitively correct when it was adopted for Scenes II, IV and V (trivially adopted in Scene I, and not adopted in Scene III) for the simple reason that each of these scenes displayed a.v.e.'s corresponding to an easily identifiable single physical object (the cube of Scene V), or a single part of one (the single straight line of Scenes II and IV). In this respect these scenes are, however, hardly representative of real life situations, which most of the time imply the presence of many physical objects possibly all moving in different ways, and this seems to require the grouping process to go right up to the level where these objects can gain their identity as individual wholes, leaving a.v.e.'s behind, and stop there, before the objects loose their individual identity in too global a v.e.. It is in this type of situation, away from the trivial a.v.e.'s and short of the single global v.e. containing all a.v.e.'s that

S-characterization can be of maximum help to Identification. Scenes VI and VII (figure 7) provide a concrete basis for discussing S-characterization.

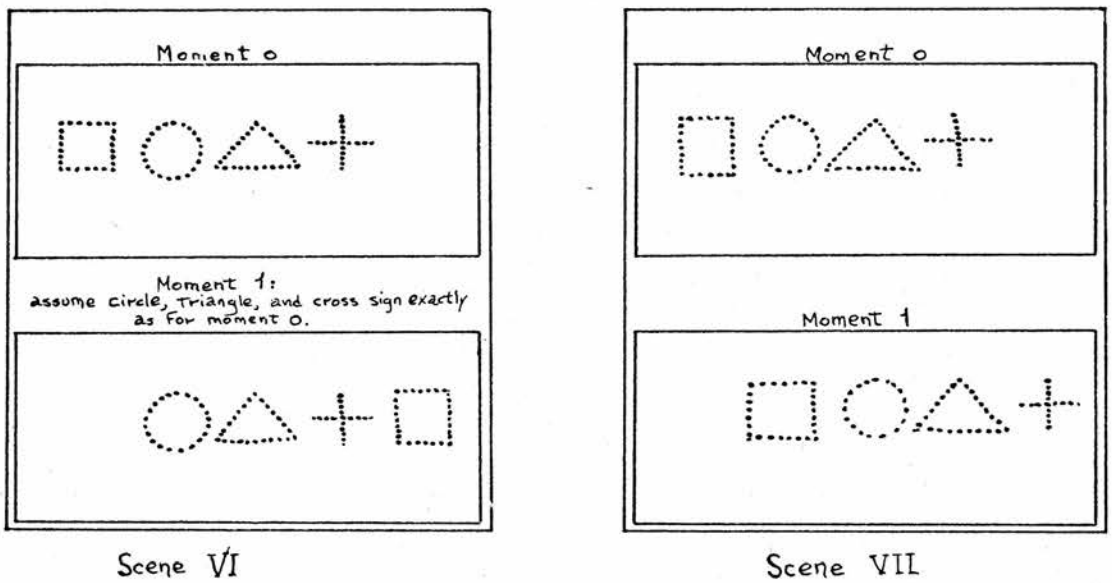


FIGURE 7. Scenes VI and VII.

The presentation of Scenes VI and VII is different from that of Scenes I to V. Here sets of a.v.e.'s detected at moment-0 and moment-1 are kept separate for more clarity, the rectangular frame within which events occur being shown "above" with a.v.e.'s detected at moment-0, and "below" with a.v.e.'s detected at moment-1. The moment-0 situation is the same in both scenes, but in Scene VI the moment-1 stimulus structure shows the square displaced from its left hand side position at moment-0 to the right hand side of the scene, while in Scene VII the moment-1 stimulus structure shows all four objects shifted to the

right. The point to be made here is that the "global" similarity of these two scenes (a "bunch of objects" on the left at moment-0, and a "bunch of objects" pushed to the right at moment-1) can be overcome by S-characterization at the level of the individual shapes to yield the desired different interpretations to the two scenes. These events can indeed be adequately distinguished and described by grouping a.v.e.'s up to the level of the four disconnected objects and by S-characterizing each object as either a square, a circle, a triangle, or a cross sign and M-characterizing each one with a global position. The square, the circle, the triangle and the cross sign being identified as such at moment-0 and moment-1, each one being given a global position, there is no difficulty in realising that in Scene VI, the circle, triangle and cross sign remain still while the square jumps from one side of the group to the other, whereas in Scene VII every object is shifted to the right by an equal amount.

It is, however, important to keep in mind that the potential diversity and the range of complexity of single-valued features (such as square, triangle, etc..) are huge, making S-characterization a potentially very expensive computational enterprise as well as a very powerful process.

This brief discussion of S-characterization more or less completes the presentation of the general problem of motion detection. Let us close this section by trying to bring the main issues together, formulating them in terms of a few critical questions. There are in fact three such questions :

First question : How much grouping is done before and how much grouping is done after computing motion? Or more precisely, given any particular stimulus event, how much grouping does the proposed visual system carry out, at any processing moment, before computing such or such a type of motion, and how much does it carry out afterwards?

Second question : How is grouping carried out? Or more precisely, at any level of processing, on which grounds is grouping achieved (i.e. is it achieved on running or frozen grounds?) ?

Third question : How is motion computed? Or more precisely, at any chosen level for motion computing, how does the proposed system identify (through time) the visual entities (or v.e.'s) represented at this level (through S-characterization), and how does the system M-characterize those same visual entities?

CHAPTER II

The state of the art

II.0 Introduction

In the past, three main scientific fields have had some involvement in trying to elucidate the problem of visual mechanisms; these fields are Psychology, Physiology, and Artificial Intelligence (A.I.). While it is true that research scientists in these fields have devoted quite considerable efforts investigating vision not more than a minute proportion of these efforts has been devoted to the "motion" aspect of vision. Furthermore, in surveying what has actually been said or shown about motion perception, if one only selects discussions which bear on possible (or even actual) processes accounting for this dimension of vision one is left with a handful of contributions which rapidly drop out of the reckoning when one asks for such processes to account for the phenomena of human visual motion detection. Nevertheless, there are quite a few relevant pieces of research which deserve to be taken into account in one way or another. Let us begin our account of these by looking at the contributions of the newest discipline to investigate vision, A.I..

II.1 Artificial Intelligence

Strictly speaking, Artificial Intelligence (hereafter A.I.) has ignored motion perception, that is to say no one in this field has yet made an explicit attempt to provide even the basis for an answer to any one of the three questions proposed at the end of the last chapter... so why bother talking about it? The main reason is that since motion perception is an integral part of the whole visual system, both depending on and taking part in the rest of the visual system's activity, any field concerned with vision is relevant to some extent, even though in this case it has mainly been concerned with "the rest of the visual system's activity".

In general terms the main concern of A.I. vision work (mostly done in the past decade) has been to "group" an input set of a.v.e.'s (sometimes sampled using a T.V. camera) on purely frozen grounds, aiming at a level of grouping where the "retinal" projection of each physical object in the scene can at worst be described as a "whole" and at best be sufficiently S-characterized to allow the desired level of recognition (e.g. this set of a.v.e.'s is a hammer, and this other set nearby is a nail which is half driven into this third set which is a log). Of course ten years of research are by no means sufficient to

lead to a complete solution (in the main only regular and simple objects in restricted contexts have been experimented on up till now) but the trend of the research is clear enough. A now classical succession of interlocking efforts along the lines sketched above is represented by the Roberts (1963) - Guzman (1968) - Huffman (1970) and Clowes (1971) - Waltz (1972) series of results. The work of Guzman on "body identification" (grouping a.v.e.'s into wholes corresponding to physical objects in a world of blocks) is a beautiful instance of designing strategies for grouping on frozen grounds up to the level where projections of physical objects are specified as wholes. However, the criteria used by Guzman to carry out his groupings fell short of being a general solution even for the very restricted universe to which he applied them, and subsequent workers modified the set of grouping criteria which he used, weaving in bits of S-characterization as stronger requirements for object identity were acknowledged (that is <u>identity for recognition</u>, or as basis for differentiation between many co-existing objects, without any reference to the motion issue ever being explicitly made). The new grouping and identification criteria were brought in from a variety of domains, naturally including the fashionable "semantics" which were said to be the main source of inspiration, and surprisingly excluding motion. Like the old ones, the new

criteria were carefully grown on absolutely frozen grounds and wisely kept single-valued (e.g. "T-junction", "front-right-vertical region", "convex edge", "on top of", etc. ...).

Now what is the relevance of all this work to motion perception? If the current task of A.I. vision work (i.e. providing strategies for adequately representing purely frozen scenes with a competence which should eventually reach the human level) is pursued without trying somewhere along the road to allow for some kind of running feature analysis, then there is only one sure way in which the resultant system could be made to allow for motion perception without requiring some demolition followed by reconstruction with new building blocks. This only way is obviously to stick motion perception at the very end of the whole frozen visual analysis, that is once everything that could be said on frozen grounds has been said. This solution which is probably intuitively favoured by most people involved in "frozen" vision can be very powerful but unfortunately fails to account for many important cases (e.g. running field effects) and unduly loads the Identification process necessary to compute motion. The "group first and compute motion afterwards" flavour of the solution is hopefully obvious, but let us specify it a little more sharply by seeing how it fits in the context

of our three critical questions about the complete problem of motion detection and how it would tend to be applied to the concrete case of the scene presented in Figure 1 (page 9).

To the first question (how much grouping before motion detection and how much grouping afterwards ?) a hypothetical defender of the solution being discussed would answer something like this : before detecting motion the visual system should carry out as much grouping as possible, the ultimate aim of this grouping being the specification of the projection of every physical object in the scene, making each one available for the ensuing motion detection. If we use frames 6 and 7 of Figure 1 (page 9) to illustrate this standpoint the task is seen as one of grouping a.v.e.'s, in each one of the two frames, up to the level where the system has specified for instance eight sets of a.v.e.'s (8 wholes) : a set for the lorry (which in fact can be specified through half a dozen sub-sets), a set for the road, a set for the hills, three sets for the three trees, one for the partially occluded sun, and one for the cloud.

To the second question (on which grounds will the groupings be carried out?) our hypothetical defender of the "group as much as possible before computing motion"

solution would answer something like : all groupings and specifications prior to motion detection should be carried out on frozen grounds, and groupings and specifications after motion detection could be made on frozen and/or running grounds. Looking again at frames 6 and 7 of figure 1, this means that the eight sets of a.v.e.'s in each frame will be defined and specified (or characterized) using exclusively frozen features (about which the A.I. vision work has a lot to say) before any running analysis can take place. However, after the motion of these eight sets has been computed new groupings and specifications can be made either on frozen grounds again or on running grounds : for instance, the identical velocities of hills and trees can serve as the (running) criterion for grouping them in a single (more global) set.

The answer to the third and final question (how will the system go about computing motion as such, i.e. how will the Identification and M-characterization problems be solved?) would first of all stress the ease with which Identification could be carried out (all objects being neatly circumscribed and S-characterized) and would go on to present some scheme for M-characterization, proposing features like global position (for each object of course) and other appropriately chosen multi-valued features covering the possible motions of the objects. Solving

M-characterization is then the only worry (which by no means implies that it is a minor one) entailed by the "group first and compute motion afterwards" solution. Looking at Figure 1 once more shows that once the system has found the lorry, the road, the hills, the trees, and the cloud in both frames 6 and 7, Identification having been readily made, M-characterizing each "object" with a global position and a global orientation which get particular values in each frame is all the system has to do to detect the respective motions of all physical objects involved in the events portrayed.

However, this solution has got at least three important drawbacks : (1) it calls for a very demanding (frozen) analysis of each frame at every moment (this is the price that has to be paid for the ease with which Identification is carried out), so demanding in fact that it is hard to see how it could be achieved as quickly as required by a temporal sampling adjusted to provide the successive inputs necessary to detect even relatively slow movements; (2) it implicitly rejects the potentially very helpful running features as part of available grouping criteria at a level prior to physical object recognition; and finally (3) if it turns out not to be realisable it will very likely require some re-thinking of the frozen analysis itself, thereby transforming an already painful

compromised progress into an even more painful regression to problems which were supposedly solved.

II.2 Physiology

Physiology, having really started providing evidence for neuronal mechanisms of vision in the early fifties, has enjoyed steady progress over one more decade than A.I., so one is not surprised to find that it has a richer literature. The first point to be made about Physiology is that its epistemological concern, instead of being centered on the "possible", is centered on the "actual"; that is to say, Physiology (and, for that matter, Psychology) is primarily interested in pinpointing the exact way in which information is handled by particular living organisms to achieve certain computational tasks. The main method used to investigate the physiology of vision, in quite a wide range of animal species, has been "single-cell recording". Very generally speaking, the idea seems to be to find, within some animal's nervous system, single cells (neurons) which respond specifically to certain characteristics of a stimulus presented to the investigated animal's eye(s). The particular specificity of such pinpointed cells together with information concerning the physiological (or anatomical) structures within which they are found (e.g. retina, geniculate bodies, different areas of the visual cortex, etc.) as well as concerning the particular animals which are used yield a basis on which the actual mechanisms responsible

for the achieved specificities can be discussed. These discussions have not however gone very far yet, and as one goes up in the evolutionary ladder fewer and fewer hypotheses are available as to what the visual mechanisms of the organism concerned could look like. As far as man is concerned, the reluctance felt by most physiologists to treat him as they treat other living organisms has saved him from the scalpel, so hypotheses regarding the physiological structure of human visual mechanims seem to be mostly based on generalisations from findings in "lower level" animals. So what has physiology to offer concerning motion detection at our level of interest?

In an interesting document reviewing the main physiological findings about movement detection Grusser and Grusser-Cornehls (1973) talk of five types of movement detectors : (I) M-neurons, which respond to stimuli undergoing translatory movement, independent of the direction of the movement; (II) DS-neurons, "for which neuronal activation mainly depends on the direction of the moving stimulus" (i.e. Direction Specific neurones); (III) OS-neurons, whose response to a moving stimulus depends on the "spatial orientation of the contrast borders of the stimulus relative to the axes of the visual field" (i.e. Orientation Specific neurones); (IV) CM-neurons, that is complex movement detecting neurons

which show specificity to particular patterns (in motion) and other (specie particular) oddities; and finally (V) Z-neurons, which respond to movement along the Z-axis (i.e. perpendicular to the retinal plane).

What does this tell us? First it tells us that some level of running feature analysis is achieved in the form of translatory motion detection (in the retinal plane with M-neurons and more specifically with DS-neurons, in the perpendicular plane with Z-neurons), that furthermore orientation, a frozen feature, is sometimes used as criterion to "modulate" the detected translatory motion of the stimulus (the exact role of these OS-neurons in motion perception being rather unclear but being interestingly interpreted by Grusser and Grusser-Cornehls as being a role derived from a "functional adaptation to the continuously moving retinal image of the stationary world" (page 411)), and that finally more complex frozen features than orientation ("complex contours", "size", and the like) are also sometimes found to be coupled with the detection of translatory motion undergone by the stimulus bearing these features (e.g. Lettvin et al, 1959).

Now what does this <u>not</u> tell us ? Mainly, it does not tell us about how those specific decisions (represented by the firing of the single cells) <u>are reached</u>, not any more than

it tells us about how those decisions are used in the overall visual system (i.e. we do not know how these decisions are taken and what is done with them). To be fair we should say that certain mechanisms have been proposed to account, for instance, for DS-neruons' decisions (cf. Grusser and Grusser-Cornehls, 1973, page 390), but so far the proposed mechanisms have remained at an extremely low level, giving some hints at how a frog or a ground squirrel might be "wired up" to carry out the computation of "directional translation of a spot of light", but failing completely to facilitate or even only allow higher level uses of the decisions which they are made to take. This is a difficult problem which has been made even more obscure by a) the extremely large range of investigated animal species (from turtle to monkey) for each of which investigators have looked for just about the same types of specific detectors and b) the extremely large range of investigated anatomical structures in particular species (from retina to deep cortex) where more or less the same types of specific detectors have also been looked for. For instance DS-neurons have been looked for and found on the one hand in the turtle, the goldfish, the salamander, the frog, the pigeon, the opussum, ..., the rabbit, the cat, and the monkey, and on the other hand in the cat's retina, optic tectum, praetectal area, lateral geniculate bodies, secondary and tertiary visual

cortex, middle and lateral supra-sylvian gyrus, and others (from Grusser and Grusser-Cornehls, 1973). The problem can be stated as follows : even though M-neurons have been found in both the frog's retina and the monkey's infero-temporal cortex, there seems to be an extremely weak probability that they lie, in the frog's eye as well as in the monkey's head, at the end of the same decision taking process (starting from the retinas' single receptors) and in the context of the same set of other decision taking processes to which their output is compared and taken even higher up for instance into motor control areas or even into conscious experience. This is really the crux of the matter for someone who is interested in designing motion detection processes which should eventually account for visual phenomena of a human type: primitive processes have to fulfil totally different requirements depending on the level of sophistication which happens to be aimed at. In other words, we believe that a visual system of a human type can hardly be thought of as being a frog's visual system with some extensions. But what about the monkey, which is surely more likely to give some more interesting hints? Evidence about the monkey's visual system has not yet reached a level where one can claim on physiological grounds that a monkey sees significantly more sophisticated things than a frog (motionwise that is); more or less the same five types of

motion detectors, and only them, have been found in both the frog and the monkey, at different levels of course but since the exact role of the level on the specific detectors is unknown we are not any wiser.

This does not imply that physiological findings should be totally discarded: it only means that answers to our three critical questions at our level of interest cannot be found directly in present day physiological data (or in available extrapolations on their basis). Great care must therefore be taken by possible model makers because the low level of motion analysis found in frogs for instance (i.e. motion detection at the retinal level) can easily guide (not to say "force") people on the track of the "compute motion first and group afterwards" type of solution. This is not necessarily a "bad" solution, but one has to weigh its pros and cons carefully before proposing it as a basis for discussing a human level of motion perception. Such a solution has for instance been adopted, among others, by Schouten (1964) in the design of a model meant to account for certain illusions of movement. Schouten acknowledges the direct influence of physiological findings on his model which in fact could be taken as a kind of model of the mechanisms behind M-neurons and DS-neurons. The important point about Schouten's attitude, however, lies not in his attempt to

model some movement detectors, but rather in the hope which he has that these receptors can be the basis of a system eventually providing a full account of human visual motion perception. Schouten indeed argues:

> "In general it seems highly promising to reconsider all known phenomena of perception of movement in terms of movement detectors". (page 55)

A quick look at Schouten's model in the light of this quote is sufficient to make one realise that what he proposes is a "motion first and the rest afterwards" type of solution. For reasons discussed in Section I.2, this solution pushed to the extreme seems incredibly risky and at any rate would require an infinitely more complex system than the simple "translation of a.v.e's" detection scheme proposed in Schouten's paper.

But nevertheless this solution is possibly applicable, so let us see what kinds of answers it would tend to make to our three critical questions. To the first one (how much grouping before and after motion detection?) the answer would be something like : first compute the translatory motions (the only possible ones at this level) of a.v.e.'s as such, or maybe wait for them to be grouped according to some field geometry (grouping on frozen grounds, under the criterion of "belonging to a given retinal region") before doing it, and group afterwards. To the second question

(on which grounds should groupings be carried out?) the answer would be that if any grouping is to be done before computing motion (e.g. field grouping), this should be done on frozen grounds, and that after motion has been computed groupings can either be made on running or frozen grounds depending on what is to be achieved. Finally the third question (how should motion be computed?) would be implicitly answered by proposing as Identification strategy the "proximity in space through time" criterion, the seemingly only available Identification strategy at the low level where motion is to be computed (where S-characterization is minimal), and by proposing position as the main feature for M-characterization (thereby allowing translatory movements to be worked out); other types of motion will of course have to be derived from these atomic translations.

All in all, we can conclude this brief discussion of physiological contributions by saying that :

1. As far as physiologists are concerned, the understanding of visual mechanisms in animals is making significant progress, having reached quite a reasonable level in the case of "lower" animals like the frog.

2. As far as people interested in "possible visual

mechanisms" are concerned, existing physiological data can be a powerful source of ideas, but great care has to be taken not to be drowned in the specificity of their context.

3. Concerning our three questions at our level of interest, physiologists leave them without an answer, but assuming direct relevance of existing physiological data and models to models of motion detection at a human level of performance leads directly to backing up quite strongly the "compute motion first and do the rest afterwards" type of solution, although no one has ever pushed this solution further than its "compute motion first..." part.

II.3 Psychology

Whereas Physiology looks for evidence of specific information available to the investigated organism in the "behaviour" of certain parts of this organism's nervous system itself, Psychology looks for this evidence in the more overt behaviour of the investigated organism (e.g. human verbal accounts of the phenomenal aspect of visually presented stimuli or this phenomenal aspect itself, reaction times to visual clues, choice behaviours involving visual discrimination tasks, etc.). A secondary difference between the two fields, but a rather crucial one, is that while Physiology has done very little work on the human visual system, Psychology has given a dominant place to this system in its investigations. But apart from these differences, Physiology and Psychology appear to be following very similar courses, both being concerned with actual organisms (often the same actual organisms), and especially with the investigated organisms' specific responses to various visual stimuli. This latter concern is responsible for the bulk of physiological and psychological data accumulated to date.

Experimental psychologists have been preoccupied with the nature and range of visual specificities for just about a century, and they have carried through thousands of

parametric investigations, bringing to light a surprising variety of phenomena, especially as far as the human visual system is concerned. Now since Psychology like Physiology emphasises the recording of specificities, it is also subject to the criticism that isolated specificities, although undoubtedly a valuable starting point, are not themselves statements of underlying mechanisms. This point can be felt most acutely by trying to derive some well known system's functional structure on the basis of its behavioural specificities alone. We tried to do it on the particular visual motion detection system which is the main subject of this dissertation and we rapidly realised how effectively isolated specificities hide the highly homogeneous and simple functional structure of the system behind their diversity and apparent complexity. Of course this does not mean that behavioural specificities are hopeless as an entry into the realm of processes, it only means that at least as much effort should be devoted to establish functional relations between specificities as to establish the nature and range of those specificities if one hopes to achieve an understanding of visual processes.

Now the important question at this point is: has anything been done in Psychology to relate recorded specificities into coherent functional systems where the underlying

mechanisms of vision at a human level of sophistication are the primary concern? Although we could not answer this question in a strictly negative way, an affirmative answer can at best be based on only a few contributions, most of which being either too local to mean anything in our context or too old to present any interest other than a historical one. This relative absence of process models to link together apparently unrelated observed specificities is, we believe, responsible for the fact, every year painfully experienced by students being taken through the history of discovered human visual specificities, that from the 1950's onwards the heavier and heavier harvest of specificities seemed to increasingly blur rather than focus the already foggy understanding of what the underlying mechanisms of human vision could look like. In fact the relative absence of process models is only partly responsible for this state of affair, a substantial part of the responsibility resting on the failure to recognise already achieved modelling results as well as a failure to follow up precise warnings expressed long before the 1950's. Max Wetheimer, in "an address before the Kant Society" in Berlin in 1924, was indeed saying :

> "There is another difficulty that may be illustrated by the following example. Suppose a mathematician shows you a proposition and you begin to "classify" it. This proposition, you say, is of such and such a type, belongs in this or that historical category, and so on. Is that

how the mathematician works ?

> "Why, you haven't grasped the thing at all", the mathematician will exclaim. "See here, this formula is not an independent, closed fact that can be dealt with for itself alone. You must see its dynamic functional (his italics) relationship to the whole from which it was lifted or you will never understand it."" (Wertheimer, 1924)

We are obviously here in the context of the Gestalt School of thought which unfortunately, against the example it was setting itself, generated in the long run more experimental than theoretical fever, and did it so well that from the 1950's onwards references to the theory itself are just about exclusively experimental invalidations or confirmations (fewer of these obviously) of the gestaltists' theoretical concepts without any attempt to alter, re-model, or replace the theory in the light of the new facts. Already in the 1950's Gestalt ideas were considered by many as a mere historical curiosity although until the uprising of A.I. a decade ago they formed just about the only available framework in which observed visual specificities could be "functionalised" . And even nowadays, since A.I. has said virtually nothing concerning motion perception, Gestalt models are still just about the only reference. So what did the Gestalt theorists say about vision, and more precisely about motion detection, that could help us answer our three questions ?

As mentioned above, the interesting point about Gestalt theorists is that they were concerned with visual processes, i.e. they were concerned with functional relationships between visually detected specificities as much as with the specificities themselves. Furthermore, they were probably considering these processes as being essentially concerned with grouping (a concept which can be found all over the Gestalt literature) an input set of atomic visual entities (a.v.e.'s), which they called "a mosaic of local sensations", into higher level entities which they called "wholes", specified through global features (which we believe is the meaning, in our terminology, of their allusions to properties of the "whole" as such which cannot be found in the individual elements alone). This interpretation of the views of Gestalt theorists can at least be felt in the following excerpts from Gestalt papers.

> "The fundamental "formula" of Gestalt theory might be expressed in this way: There are wholes, the behaviour of which is not determined by that of their individual elements, but where the part-processes are themselves determined by the intrinsic nature of the whole. It is the hope of Gestalt theory to determine the nature of such wholes". (Wertheimer, 1924)

> "Our view will be that, instead of reacting to local stimuli by local and mutually independent events, the organism responds to the pattern (his italics) of stimuli to which it is exposed; and that this answer is a unitary process, a functional whole, which gives, in experience, a sensory scene rather than a mosaic of local sensations". (Kohler, 1947)

The concern for groupings within the visual system might have led to some answers to our three questions at our level of interest, but for two main reasons such answers were never provided. The first reason is that, as hinted at by Kohler (above), the main preoccupation of early Gestalt psychologists was to convince comtemporary psychologists of the need for grouping more than to elucidate the problem of how the grouping(s) should be carried out. The sad thing is that Gestalt ideas were "put on the shelf" before this point was made successfully, so that in the end very little energy was spent on discussing grouping processes themselves. The flavour of the polemic about the need for grouping can be felt in the following statement by Kohler:

> "If it could be shown that all elementary segments of a sensory surface and the corresponding central field were physically absolutely insulated, then we should be justified in concluding that there were as many physical systems and hence local processes as there are individual pathways. But where is the evidence for such an assumption? Indeed we find abundant evidence for the opposite point of view, e.g. between two unequally stimulated parts of the retina there is an electro-motive force, but this would be impossible without functional (osmotic) communication between the two regions ... Since there is a multiplicity of such transverse functional connections between successive niveaus of sensory sectors, the histological reasons given for studying chains of neurones as physically independent systems are without foundation". (Kohler, 1920)

Of course, in order to make their point about the need for

grouping, the Gestalt psychologists had to show that certain facts of vision could only be accounted for by "grouping", which implied that they should at least hint at ways in which some groupings could be carried out if they wanted to have any impact at all. This they did to a certain extent, but the way in which they did it brings us to the second reason accounting for the absence of answers to our questions: they used as process models for visual groupings contemporary notions about essentially "physical" processes in such fields as magnetism and electricity, fluid dynamics, chemistry, etc.. Even though this was the most sensible thing to do at a time when few other means of modelling were available this use of physical concepts rapidly reached exhaustion without having provided sufficiently powerful and detailed grouping schemes. In other words, although the physical models proposed did account for certain specificities they did not provide any means of discussing them in terms of what their respective roles and their possible inter-relationships could be in the overall visual information processing system. The physical model used could hardly provide a basis for discussing if such or such a detected visual feature is used as "grouping criterion" or as "incidental characterizing feature", or if such or such a feature belongs to the computational history of some other feature, or if such or such a

feature is running or frozen. This does not mean that Gestalt psychologists had not implicitly recognised that groupings could occur through time (running features) as well as through space (frozen features), nor does it mean that they had not noticed in some way or other that criteria for grouping could be either frozen or running; it only means that their models could very hardly provide an explicit framework within which these distinctions could be handled and articulated properly. This can be felt quite strongly in the following excerpt from a translation of Kohler's 1920 paper on "physical gestalten". The excerpt has been extracted from the context of a discussion of Wertheimer's famous Law of "Pragnanz" governing grouping processes in the "whole optic sector".

> "A very different example shows that we are not dealing here with peculiarities of electromagnetic and electrodynamic processes, but with general properties of nature. The illustration is taken from van der Mennsbugghe. A soap film is enclosed by a plane frame of wire and a small loop of very fine thread is placed in an irregular form upon it. If one pricks the film inside the loop, this part of the film vanishes, and the thread is exposed only to the surface tension of the outer film. These forces tend to give to the region enclosed by the thread the largest possible area, so that the remaining film has the smallest possible area. The thread thus immediately becomes a circle. Where a physical form of homogeneous material properties can yield sufficiently to the systemic forces acting upon it, it seems to be a general rule that the very simple and regular spatial arrangements are reached in a stationary state." (Kohler, 1920)

One can readily see from this quote a brilliant example of the essentially "continuous" nature of Gestalt's analogue models: no discrete decisions are explicitly represented in their "field dynamics"; everything seems to be happening all at once through complex inter- and counter-actions of flow lines of energy. Computational precedences are therefore lost to the model user, since although they can be argued to be implicitly taken into account in the overall process they are not explicitly available for argument. This weakens the model considerably, and accounts for Gestalt's complete failure to provide even the beginning of an answer to our first question (at which stage of grouping should motion be computed?).

Concerning the second question (on which grounds should we group?), although we lack an answer to the first question about the level at which grouping should be carried out, we can at least find in some observations made by the Gestalt psychologists an acknowledgement of the very important fact that the human visual system does use running features as well as frozen features as criteria for grouping (at some level or other). This means that in the human visual system running features are not exclusively "ends in themselves" but are also in some cases directly involved in the grouping process as such.

Quite apart from the learning issue stressed in the following quote, one can see quite clearly in it the acknowledgement, by both Kohler and Wertheimer, of at least some running features being used as grouping criterion by the human visual system.

> "In one form of empiristic explanation it is said that we have learned to regard as wholes whatever always moves together. Wertheimer has pointed out that, if some parts of the field begin to move at the same time and in a uniform way, they become at once a moving unit. In other words, if a "common fate" actually determines sensory grouping, it does so as a factor or primary sensory organisation rather than via processes of learning". (Kohler, 1947)

To the third and last question (how should motion be computed?) the Gestalt psychologists have also provided some fragmentary answers, but lacked a sufficiently detailed discussion to cover the more precise issues of Identification (and consequently S-characterization) and M-characterization. This of course is partly accounted for and partly accounts for the lack of precise answers to the first question. The Gestalt psychologists' more precise ideas about motion detection were embodied in a model mainly accounting for what they called "β-movement", that is apparent translatory movement of any "whole". This model is probably the best known of all the Gestalt models and is another rather good example of their essentially "physical" nature. The model can be understood rather easily from Hartmann's description of it

in the following excerpt from a translation of his 1923 paper.

> "What, now, is the physiological process corresponding to the phenomenal experience of β-movement? If a retinal area is stimulated by incoming light, the energy density of this region increases until a certain constant value is reached; the dynamic current becomes a stationary Gestalt. When the source of light is withdrawn, the energy density decreases - but more slowly than it arose. If, now, another area is at this moment increasing in energy density, an exchange of energy between the two will take place and the latter will thus be influenced by the former. Designating the first as I and the second as II we may describe the result by saying that the distance between I and II will decrease - i.e. the current passes from I to II. The phenomenal correlate of this is an experience of β-movement.
> An analogous phenomenon in hydrodynamics will illustrate this passing of a current ... . In the bottom of a large water container there are two round holes. The water is shallow. At first both holes are closed. When one of them is opened the place where the water is sucked down can be readily identified by a little whirlpool upon the water's surface. If both holes are closed and one is opened for a brief time, closed, and the other opened after a short but definite interval, the transmission from one whirlpool to the other can be easily observed. The current passes from I to II. In the case of a β-process, the current passes from brain area I across an unstimulated region to brain area II". (Hartman, 1923)

The main problem with the model is that first of all it explicitly restricts M-characterization to position, limiting explicit motion detection to translatory motion only, the remaining aspects of movement being handled by much looser concepts which do not deserve to be called models. However, other types of movement can only exist

in the case of a "whole" (or v.e.) which belongs to a higher level than a.v.e.'s (these "point-like" entities having only a position to offer as interesting M-characterizing feature). Does this mean that in the case of a.v.e.'s the Gestalt model is adequate? Not quite, since in many cases the model will run into the Identification problems discussed in the context of the "compute motion first solution" (see Section I.2). For instance try the "plug hole" scheme on Scene V, Fig. 4. This brings us to the second weakness of the model, which is its clumsiness in handling more than one "whole", Identification problems in these cases having been totally overlooked. This might have been due to the very restricted and simple types of stimuli used by the early experimenters. The "identity" problem was however taken up by Ternus (1926), but great care seems to have been taken to discuss the resulting observations outside the scope of the β-movement model or any other explicit model for that matter. Ternus' very interesting observations were made using for instance a pattern of dots in which some dots were made to occupy more or less the same retinal loci from one moment to the next while the rest of the pattern's dots would change place: he studied the conditions under which the "still-dots" would "loose their identity" as they are being "dragged" into a motion where the identification criterion seems to belong to each dot's

relation to the whole more than to its retinal locus. From Ternus' discussions it is however extremely hard to see if he means that each dot's identity is determined after the whole has been assessed globally, each dot then moving accordingly, or if motion is computed on the whole as such then giving to each dot an identity on this basis. Anyway, in neither case does he propose a model.

From about the time of Ternus' paper, we can already say that the Gestalt explicit model of motion perception starts being overtaken by the results of experimental investigation, and although the general Gestalt principles remain vey influential most investigators turned their attention to behavioural specificities backed up with very primitive models. After the example of Ternus, who showed that wholistic properties of sets of elements were determinant in deciding on the identity of local elements through time, we have the brilliant example of Duncker who, in the late 1920's, set out to explore the actual dependence of motion as such on the wholistic properties of the stimulus, showing that the motion of visual objects relative to their visual context dominates their motion relative to the retina in human visual processing, but like Ternus steering well clear of any kind of explicit integration of this principle in the modelling context of electrodynamics.

Indeed, although the argued dependence of motion on relational properties of elements of the stimulus fits perfectly in the global Gestalt framework, its main requirements - that relational properties be worked out first and then be made to affect motion detection in the desired way - reach far beyond the explicit motion detection scheme modelled through Gestalt traditional electrodynamics concepts (cf. the "plug hole" analogy). To provide some ground on which the above abstract arguments can rest it is probably sufficient to describe the best known experiment carried out by Duncker who used a setup where a small (stationary) spot of light (2 cms. in diameter) was projected onto a piece of cardboard (66 x 48 cms.); subjects stood 1 m. away from the stimulus. "When the cardboard was moved back and forth, the fixated spot of light appeared to move also but in the opposite direction" (Duncker, 1929). On the basis of this and other observations Duncker formulated his Law of Motion Distribution:

> "The phenomenal motions of separating objects is determined by the kind and degree of mutual "localisation" of these objects, and this whether one of them is localised relative to the other or both are localised with regard to each other respectively. In other words : phenomenal motion is displacement in a natural frame of reference". (Duncker, 1929)

The interesting point here is that Duncker is refusing to let "position <u>relative to the retina</u>" (like that of

a.v.e.'s) be the sole dimension of M-characterization, hinting at least at "position relative to the natural frame of reference" as an important aspect of M-characterization. However, the criteria under which sets of a.v.e.'s should be analysed to show for "kinds" and "degrees" of mutual localisation (cf. last quote) are, to say the least, rather fuzzy (apart maybe from the strict "topological enclosure") and the exact way in which these criteria should determine motion detection is totally overlooked.

Concerns such as Duncker's were taken up to form the basis of a very rich vein of experimental investigation which extends at least from 1950 to the present time. We are referring here to the work of the Uppsala School under Gunnar Johansson whose first investigations on mutual interactions of visually perceived moving elements can be found as early as 1950 in his book Configurations in Event Perception (Johansson, 1950). However, our interest in modelling has to wait until the 1970's before finding explicit reference to a process model in Johansson's writings. The model is based on the use of classical mathematical concepts such as vector analysis and projective geometry; the main idea behind the model is as follows :

> "We will find that the pattern of change hitting the retina is analysed perceptually in units of

> common motion states and motion relative to these components. The motion of a given object or element can perceptually take part in two different motions at the same time. The motion component which it has in common with other simultaneous displacements in the visual field forms one separate motion and the deviating displacement relative to this common component constitutes another." (Johansson, 1973)

Johansson's attitude seems to depart a bit from Duncker's by not stating as a basic principle that the basic motions from which vectorial components are extracted are themselves subject to relative representation; we believe in fact that Johansson's attitude towards basic motion detection is to compute translations of a.v.e.'s (or something close to this level) relative to the retina and then carry out groupings on the basis of the vectorial components of these translations. Whereas in Duncker's "law of motion distribution" the starting point is (global) localisation, Johansson's argument, although it is expressed in a somewhat more formal language, starts from motions. This distinction can be more clearly grasped by realising how badly Johansson's model accounts for induced motions of the type presented above (e.g. the cardboard-spot—of—light experiment) in contrast with how well it accounts for the "two bulbs on a wheel" experiment where a wheel with a light bulb on its hub and one somewhere on its periphery is set into motion in the dark along a straight line in a fronto-parallel plane relative to the observer. When only the "periphery" bulb is lit a

"hopping" motion of the light is seen (Figure 8a) whereas when both bulbs are lit the "periphery" light is seen as revolving around the "hub" light while both are seen as translating together along the wheel's path (Figure 8b).

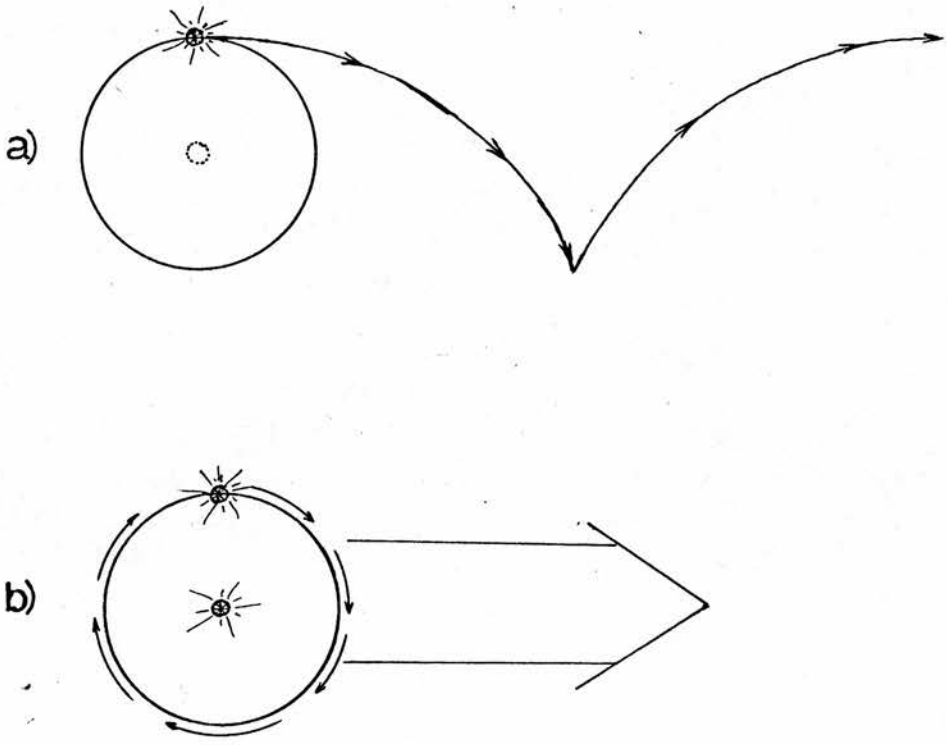


FIGURE 8. Hopping/revolving light bulb experiment.

Johansson's scheme really seems to have been worked out for cases where a motion can be reduced by having one of its components taken away (e.g. the translatory movement taken from the peripheral bulb on the wheel) but not for cases where a motion can be increased by adding to it some velocity component (e.g. the spot of light having the converse of the cardboard's velocity added to its own). If this is what he wanted to do, this is fair enough, but the vectorial components model, although more detailed, does not cover the range of Duncker's law. We believe the most interesting feature of the vectorial comoonents model to be the clearly expressed grouping strategies of the moving elements. Although these strategies can be fairly well understood from the last quote, the following one can leave no doubt regarding their spirit :

> "When in the motions of a set of proximal elements equal, simultaneous motion vectors can be mathematically abstracted (according to some simple rules), these components are perceptually isolated and perceived as one unitary motion."
>
> (Johansson, 1971 b)

This grouping on the basis of common velocity components is we believe a beautiful instance of a "common fate" determining sensory grouping (cf. Kohler's comment reported on page 78). The descriptive and predictive power of the velocity components model reaches its peak in a very impressive demonstration of a man wearing a light bulb on each side of each of his main limb joints and

filmed in such a way that only the lights can be seen when the film is projected. If the man is still (say sitting on a chair), only a random set of dots is seen, but as soon as he starts moving (say getting up and walking), the lights are immediately grouped in a way which unmistakably specifies a man. The power of running features as grouping criteria is in this case hard to challenge, but the exact way in which the hierarchies of velocity components are worked out by the visual system remains unclear.

It might be worth digressing here since there is even more powerful evidence than the "walking man" that the human visual system does use running features as grouping criteria. This evidence can be found, discussed in different ways, in quite a few recent writings (e.g. Julesz (1971), Kaplan (1969), Lee (1971, 1972)) and bears on the extensive use by the human visual system of running features as grouping criteria in the absence of the normally coexisting frozen criteria which can just about always be argued to be responsible for most groupings. The situations investigated by the above mentioned authors put the emphasis even more on running structures by working with strictly random frozen patterns (in the spirit of our Scene III, figure 4) where the "running patterns" are designed to be coherent enough to

overwhelmingly reduce the uncertainty of groupings ... if the observing visual system has got access to running features as criteria for grouping. The results are unanimous : the human visual system does have this power of grouping on running grounds, whether or not frozen constraints bearing on the same groupings are present. The following excerpt from Julesz' Foundations of Cyclopean Perception (Julesz, 1971) expresses well the point being made here, and interestingly echoes Max Wertheimer (cf. page 78):

> "As I discussed, clusters of dots sharing some common properties are the basis of object separation. One of the most important properties underlying the grouping of many dots is movement of the same velocity and orientation. Even if dots are not adjacent, their common motion will be perceived as a rigid transparent object in motion. In 1966, Julesz and Bosche prepared a computer-generated movie ... in which a certain proportion of random black and white dots moved to the right, while the rest moved to the left. Even these non-adjacent thousands of random dots grouped together into two transparent oppositely moving surfaces. Of course, if the dots moving together are spatially adjacent, the cluster formation becomes even more pronounced." (p.107)

It is interesting to realise that the grouping criteria Johansson talks about are more elaborate and powerful than those Julesz mentions, the former talking about common components of velocities whereas the latter is talking about common velocities. However, although the "vectorial components" model accounts for the observed outcome in the

case of the patterns used in the above mentioned Julesz and Bosche's movie, it does so mainly because there is no common component between velocities of dots going right and dots going left thereby bringing to equality divergence of velocity and divergence of velocity components. If the two sets of random dots were made to move differently, say one set going straight up and the other going due left, then grouping on the grounds of common velocity would yield a different result than grouping on the grounds of common components : the former would still differenciate between two definite groups of dots, one going up and the other one going left, while the latter would group all the dots together under the criterion of "a common velocity component" combining "up motion" with "left motion" into a single 45 degree (10:30 hand direction on the clock) global motion of two separating sub-sets of dots. Although we have tried to see this global motion along the "common component" of the motions of random dots moving in such a way we have not succeeded; what could be predicted from grouping according to velocities as such was always the observed outcome: the two sets of dots remained well separated. This, of course, does not mean that the vectorial component scheme should be rejected; it only indicates that priorities might have to be taken into account when trying to model human visual processing, allowing for grouping on

"vectorial components" grounds but considering first the coarser (i.e. less flexible and less powerful, computationally speaking) evidence of velocities as such.

Let us now turn to more general criticisms of the vectorial components model. These criticisms can be best presented in the context of our three questions about motion detection schemes in general. Concerning the first question (how much grouping before and after computing motion ?) we have already expressed our belief that Johansson favours a very low level motion analysis, allowing for very little grouping to be done beforehand. This can be clearly realised from Johansson's definition of the "proximal elements" whose motions are vectorially analysed in the context of his model :

> "In our everyday life visual motion perception is generated by continuous changes in the proximal optical energy structures on the retina. The inhomogeneities in these structures can be treated as built up from elements, i.e. small optical pencils of light (stationary or moving) of rather constant brightness. Stimuli perceptually indicating real motion are built up from proximal motions in such inhomonogeneities or elements." (Johansson, 1971a)

Furthermore, this "compute motion first and group afterwards" orientation can be felt in Johansson's belief that "the recent research in sensory neurophysiology for which Hubel and Wiesel are highly representative has given the theorist in perceptual psychology a new and profitable

physiological basis ...." (Johansson, 1971a); we interpret this as a sign of belief in low level translatory motion detection of the type used for instance by Schouten (1964) in which case the reference to neurophysiological findings was also observed (cf. p.65). Finally this attitude towards our first question can be traced in the work of research workers associated with the Uppsala School at some point or other like Lee (personal communication) for instance whose recent idea of retinal motion detectors directly specific to vectorial components of elementary translatory movements can hardly be thought of outside the "compute motion first and group afterwards" type of paradigm. Of course, our point here is that although this is a perfectly legitimate attitude to adopt, the bulk of the problems lies in this case in the "group afterwards" part of the scheme.

Regarding the second question (on which grounds should the groupings be carried out?) the model rather explicitly states its extensive use of running features as grouping criteria, but the exact way in which all higher level types of movement are to be achieved on such basis remains unspecified (e.g. what is a "pendulum motion"? what is a "rotation"? what is a "transforming shape"?), and, as argued in Section I.2 when discussing Scene IV (see pp. 41-42), this is really the main problem with the "compute

motion first" solution. Still concerning the second question, but this time regarding the idea of using vectorial analyses as basis for grouping, explicit vectorial analysis is not only proposed as a way of describing what human subjects see, it is also proposed as a computational model of how the human visual system does generate the perceived outcome. This is a dangerous standpoint because it implies that it is computationally possible, given nerve nets of some kind, to carry out this type of computation within the time and space limits which are imposed on the visual system by the environment in which it lives. This problem has been felt most acutely in the now classical "flop", in the early days of computer modelling, of the exhaustive formal description of the game of Chess which, although it was theoretically (or mathematically) perfectly legitimate, turned out to be so far beyond acceptable spatio-temporal constraints on computational performance that it totally lost its interest. It is one thing to give a legitimate account of logically possible bases for decision taking in a given environment, but it is quite a different matter to discuss the modalities of the decision taking processes which use some or all of these bases to adapt to the environment. It seems more and more accepted in fact to regard descriptions of problem spaces in terms of classical mathematics as very poor indicators of what could be an

interesting or actual process which would behave adaptatively in these problem spaces. However, not everybody shares this view, the J.J. Gibson School of thought (e.g. Gibson, 1966) being a good example of the opposite one. The Gibsonian standpoint is well outlined in a paper by Lee (1974) from which the following excerpt, taken from the paper's summary, is enlightening :

> "A mathematical description has been developed of the optical flow-pattern at the eye of an observer moving along a rectilinear path through the environment, and from this has been derived some of the basic optical information that is available about the environment and about the observer's movement relative to it ...
>
> The optical flow-pattern has been described in terms of the optic velocity field on a cylindrical optic projection surface. It should be noted that this particular projection surface was chosen simply because it enabled the mathematical analysis to be expressed in a particularly simple form. However, for some purposes, it may well be more convenient to consider the optic velocity field on a spherical projection surface, as Gibson has done, or on a planar projection surface ..... But whichever projection surface is chosen for a particular analysis, the results of the present analysis, of course, still hold; the mathematical formulae simply have to be appropriately transformed ....
>
> In general, the analysis demonstrates the considerable amount of information, both about the environment and about the observer's movement relative to it, that is <u>directly available</u> in the optic velocity field at a moving observer's eye. To avail himself of this information the observer, of course, needs a visual system that is capable of registering the optic velocity field and its derivative properties. The fact alone that most animals, including man, can visually guide their locomotor behaviour would suggest that they have such a visual system. There is also more direct evidence available about the human visual system

> (e.g. Gibson, 1957; Johansson, 1950; Kaplan, 1969; Lee, 1971). But, precisely how, and to what degree, a particular organism can pick up the available visual information is an empirical question, and as such is beyond the scope of the present paper. The analysis, however, does raise a number of well-defined empirical questions that may well be worth pursuing."

Lee's interest, and also Gibson's for that matter, clearly lies in using the formalism of classical projective geometry to describe potentially detectable information in the environment within which particular organisms have to behave adaptatively. We believe that this is quite a safe enterprise for as long as one is not directly concerned with finding out "how and to what degree" (as Lee himself puts it) the available information is picked up by some organism or other. However, as soon as one turns towards these problems then the choice of the formalism used to describe the "information space" starts playing an important role in generating hypotheses to be checked experimentally and it is on the choice of this formalism that the success or failure of the enterprise mostly rests. This means that although it is of course true that strictly speaking "precisely how, and to what degree, a particular organism can pick up the available visual information is an empirical question", we believe that any empirical investigation has to be carried out within some hypothesis generating framework which in this case should describe "the available visual information" in terms which

are as compatible as possible with the level of information processing at which the investigated organism is likely to perform. The problem is of course that one does not know prior to the investigation what the investigated organism's visual analysis strategies look like, but what we want to stress is that investigators should be prepared to change their formalism as soon as its unsuitability in any given investigation is noticed. This is what happened in the case of Chess for instance. Concerning the description of the visual information available in our environment in terms of classical mathematics (e.g. projective geometry, vectorial analysis) we tend to believe that it is not suitable as theoretical framework for generating hypotheses about precise human or animal visual analysis strategies because it calls for computations which would require so much processing time and energy that we doubt if any organism would be capable of reaching the results in time to be able to use them to adapt successfully to the environment.

Finally, to the third question (how should motion be computed?) the vectorial components model offers the same answer as any model computing motion at the retinal level : Identification is done on the basis of proximity and M-characterization is limited to position, with all the problems that this entails.

Now we have to revert to the Gestalt context to deal with the last model of motion perception relevant to our problem. The model is that developed by Kolers (1972). Kolers' model is contemporary to Johansson's and, like the latter's, sets its roots deep down in the Gestalt context. However, Kolers' model is totally different from Johansson's and deals with apparently totally different aspects of motion perception : the two authors hardly refer to each other and really give the impression of working in totally different areas of Psychology. So what does Kolers talk about?

Very early in his book, Kolers makes a rather interesting distinction between two aspects of Gestalt theoretical tendencies :

> "Wertheimer and the Gestalt investigators seem to have been of two minds about the matter. The short-circuit theory makes no mention of figures and shapes; it talks about radiating patterns of excitation, in Wertheimer's (1912) version, and columns of electrical excitation, in Kohler's (1923). But then Ternus (1926), under Wertheimer's direction, extended the discovery of Gestalt organising principles (Wertheimer, 1923) to apparent motion, and when von Schiller (1933) discussed the tendency of the spatially and temporally disparate shapes to become assimilated into a Gestalt unity, figure and its organisation were clearly important principles in mind. The result is that two models can be formulated from the earlier work. One model states that only a disparity between locations of stimuli is perceived and motion is created to resolve it; a second model states that a disparity between figures in different locations is perceived, and motion is created to resolve that." (Kolers, 1972)

Although we do not agree with this dichotomy as a fair account of Gestalt ideas, it contains a rather interesting processing distinction which fits very well in our context and which provides a beautiful introduction to Kolers own ideas about the analysis of movement. In our terminology, we believe that what Kolers is saying is that one Gestalt model (the explicit "plug hole" or "short circuit" one) limits itself to carrying out Identification using what we usually refer to as the "proximity strategy", M-characterizing visual entities through position only, whereas another model (which we believe to be the implicit models of Ternus and von Schiller) carries out Identification on the basis of a much more elaborate S-characterization process (accounting for "shape"), M-characterization being there again limited to position.

Kolers goes on to show, throughout his book, that neither one of those two models is completely satisfactory although the first one seems to account for more empirical evidence than the other. This conclusion is reached on the basis of quite a number of highly interesting observations indicating that in most cases involving temporal successions of frames in which a few simple shapes change position and/or shape from moment to moment (e.g. our Scenes VI and VII, figure 7, with squares, triangles, circles, and cross signs) Identification is

rarely based on individual global (frozen) characteristics (i.e. S-characterization, e.g. shape) of the different

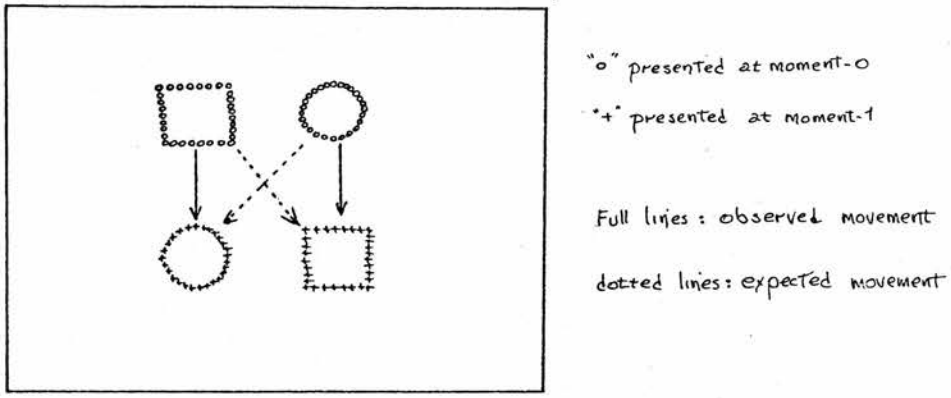


FIGURE 9. Stimulus set up used by Kolers (1972) to show that "shape" is is not taken as basis for "identification" in some motion detection situations.

objects. The most representative instance of this oft observed irrelevance of shape identification is the scene represented in Figure 9, where human subjects were never able to see motion along the dotted lines (representing the expected motion if shape was the basis of Identification): motion was always reported to be vertically downwards for both shapes, however hard the subjects tried to obtain the diagonal crossing motion. However, although Kolers found that this was a strong tendency it turned out that "shape" was not altogether ignored in all situations. According to Kolers, it seems to be the case that in situations analogous to that of our

Scenes VI and VII (figure 7) the human visual system manages to solve the problem by taking some aspect of shape into account : this allows the system to distinguish between the case where the first "object" jumps over the remaining three objects and the case where all four objects are shifted together. So since to Kolers' eyes both "Identification on the basis of proximity" and "Identification on the basis of shape" seemed to play a certain role in motion detection (although the first one seemed to carry more weight) he proposed a model that would account for this duality of processing. He describes it as follows:

> "A visual stimulus whose image falls on the eye may be thought to generate two signals. One is a spread of excitation throughout the nerve fibres of the retina itself, which will be called a Horizontal signal. The other is a message to deeper parts of the nervous system, which will be called a Vertical signal. The Horizontal or H-signal is ideally suited to represent information about the location of a stimulus. The Vertical or V-Signal is equally well-suited to represent information about identity. Thus the H-signal supplies information about where something is, and the V-signal supplies information about what it is."
>
> (Kolers, 1972)

In this model, however, duality of processing is the only thing explicitly accounted for, the dominance of "proximity" over "shape", which we believe is a most important point, being un-accounted for. Furthermore, the actual mechanisms underlying each of the two proposed

aspects, especially as to the exact nature and the generating principles of the V-signal (which contains "everything except location"), are totally skipped. As far as the H-signal is concerned, the lack of precision in the description of its horizontal journeys across the retinal network makes of it a concept which explains phenomena at least as loosely as the Gestalt concepts of "electrical spreading" or "attraction" (Kolers having said that "the concept of attraction is clumsy at best" (1972, p.71)), and by restricting it to the retinal level we believe that Kolers slips behind the Gestalt model which allowed its most basic processes to have the "whole optic sector" (from retina to deepest brain) for playground. We indeed believe that the main weakness of Kolers' H-signal lies in its extremely low level context : the retina. This can only mean that locations carried by H-signals are a.v.e.'s locations (or something very close to them), the specification of global locations of higher level wholes requiring a "computational space" which we do not believe the retina can offer. Would this mean that Kolers is advocating a "compute motion first and group afterwards" type of model for the human visual system ? Maybe not quite, but his ideas seem to lie much closer to this type of model than to its converse. To sum up Kolers' model it explicitly advocates M-characterization on the basis of <u>position</u> only (the level of this position being uncertain)

and Identification on the basis of position (i.e. proximity) and/or more powerful S-characterization (loosely defined as "shape"), no details being given about which criteria are used in what contexts, but the weaker role of "shape" being stressed outside the model itself on a rather interesting empirical basis.

This concludes our discussion of what Artificial Intelligence, Physiology, and Psychology have to offer in terms of possible or actual solutions to our problem. Nowhere have we seen even a setting of a solution which would cover, even in terms of primitive concepts, the few types of scenes used to state the problem in Section I.2. All solutions hinted at were relevant only in the context of very limited parts of our problem space, looking inapplicable or at best very clumsy in the context of the other parts. We therefore seem to have complete freedom of action: just about any path through the whole problem space is bound to be a step forward.

CHAPTER III

The proposed solution

In Chapter I, a problem was defined and three questions were formulated to crystallize the critical issues which it contained. In Chapter II, it was argued that the literature in the relevant fields of research is devoid of satisfactory answers to these questions at our level of interest. In this chapter we are seeking one set of satisfactory answers to these questions, making explicit the macro-structure of our motion detection system.

The main dilemma we are faced with in confronting such a task arises in the context of discussing the computation of the different possible movements of the physical object(s) in some observed scene, as opposed to that of discussing the role of detected motion as grouping criterion, and can be expressed as follows. On the one hand we can choose to have motion computed before having most of the grouping carried out; what is nice about the working of a system using this global strategy is that M-characterization is relatively easily done and that when the proximity strategy is successful Identification is trivially achieved. However, there are drawbacks - the

two main ones being that when the proximity strategy is fooled Identification becomes totally hopeless, and that when more global motions have to be computed on the basis of lower level ones (e.g. a rotation derived from "atomic" translations) computational complexity can reach undesirable heights. On the other hand we could instead choose to have the grouping carried out before having motion computed, but here again there are pros and cons. What is nice about this strategy is that it reduces (proportionally to the extent of the grouping) the number of visual entities whose motions have to be computed; but what is sad about it is that the grouping itself indirectly causes a loss of information, and if there is more than one moving physical object to be identified in the scene then fancy grouping criteria and S-characterization schemes will have to be carried out before motion as such can be computed.

After having weighed these two alternatives it was decided that the "group first" strategy had fewer and less serious drawbacks than the other strategy, and that we should cautiously start moving in this direction, trying as far as possible to steer clear of those aspects of the solution which are responsible for its two main drawbacks, namely the loss of information and the complex grouping and S-characterization. These two drawbacks were tackled

in the following way.

Loss of information arises from the fact that the visual entity (v.e.) resulting from a grouping can hardly be characterized through global features specifying this v.e. in a way which conveys all the potential diversity of the characterization of all its local elements. But of course the importance of the losses through grouping can be minimised by choosing the most appropriate set of characterizing global features; what we should therefore try to achieve, as far as motion detection is concerned, is an M-characterization through a set of multi-valued features conveying all the "relevant information" existing potentially at the level of the more numerous and more local v.e.'s grouped into the single v.e. to which this M-characterization is to be applied. What we are proposing here, given for instance a situation where many a.v.e.'s have to be grouped into a single more global v.e., is to compensate for the decrease in number of the many local v.e.'s (i.e. a.v.e.'s) M-characterized through but a single feature (namely retinal position) by increasing the number of M-characterizing features at the level of the single global v.e.. These global M-characterizing features should be sufficient to provide the system with a basis for computing any motion bearing some significance in the context of the behaviour of the

organism which this system serves (in our case of course we are talking of a human level of visual needs). The choice of actual global M-characterizing features of this kind and their fitting in a precise working system are discussed in detail in Part II.

However, even given the adequate global M-characterizing features required by the scheme discussed in the above paragraph, the system's task will only be made nice and easy in cases where the scene to be analysed consists of a single moving physical object. In this case only will grouping criteria and S-characterization strategies be easily handled, both of them trivially (but adequately) resting on the "existence" of the v.e.'s they are concerned with. This brings us to the second drawback to be eliminated, which is that as soon as more than one physical object is implied in the moving scene not only is there a problem in deciding on which criteria to group and how to stop grouping when the level of the physical objects is reached, but there is also the problem of adequately S-characterizing each object-v.e. for Identification purposes in the motion detection scheme itself. Furthermore, the fact that there are many objects to be detected implies as many M-characterizations as there are objects and thereby increases accordingly the number of motions to be computed (the final number of

motions to be computed being determined by the number of M-characterizing features times the number of v.e.'s yielded by the grouping process).

To avoid the problems, we have to design our system in such a way that any set of a.v.e.'s "looked at" by the system is made to yield, through grouping, a single v.e. on which motion will be computed (on the basis of the changing values of each one of the global M-characterizing features made to bear on it). Now in order to allow for this single v.e., which we will from now on call the visual object, to be represented or described in a sufficiently powerful and flexible way we decided to make the system provide, out of the set of a.v.e's being looked at and besides the visual object itself, two other "groups" of a.v.e.'s. One of these two "secondary groups" will be totally separate from the visual object's group of a.v.e.'s (i.e. one secondary group of a.v.e.'s and the visual object's group of a.v.e.'s will be disconnected sets of a.v.e.'s) and will in fact consist of all a.v.e.'s "looked at" but not chosen to be part of the visual object : this group of a.v.e.'s will be considered as being the visual object's outside frame of reference, and will from now on be called the background. The other "secondary group" of a.v.e.'s will provide the visual object's inside reference, and will therefore be a subset of the visual

object's set of a.v.e.'s; it will be used to describe the "shape" of the visual object and will be called the visual sub-object. These three sets of a.v.e.'s will be used to derive the visual object's values for all M-characterizing features specifying it, and it is on this basis that motion will be computed. This single-object scheme means among other things that quite often a single moving visual object might well consist of the projection of many physical objects, and at some other time consist of the projection of a part of one physical object, the system's power lying in its ability to shift the status of parts of scenes from "background" to "visual object" to "visual sub-object" (and vice versa) in order to analyse them from all sides and in all sorts of inter-relations. But this can only be achieved if the system has got some flexibility regarding the choice of the a.v.e.'s which it will "look at" and irrevocably (for a given moment) sort into the three definite sets. A certain flexibility can of course be allowed by making the system's retina move (with the eye) in the environment and change the input set of a.v.e.'s by scanning this environment, but this is hardly enough to allow the system to exploit the full potential of its three-sided grouping strategy. What one wants, besides a freedom to "translate" through the scene, is freedom to expand or contract the field of view, or maybe even freedom to select on the basis of higher level

features. This is why the concept of an "attentional retina" was created. This attentional retina is loaded from the ordinary or "physical" retina according to some criterion and it is the content of this attentional retina which is split into our three crucial sets of a.v.e.'s. So by expanding this attentional retina, contracting it, and moving it about within the physical retina which can itself move about the environment the system has access to any part of this environment and any one of its elements.

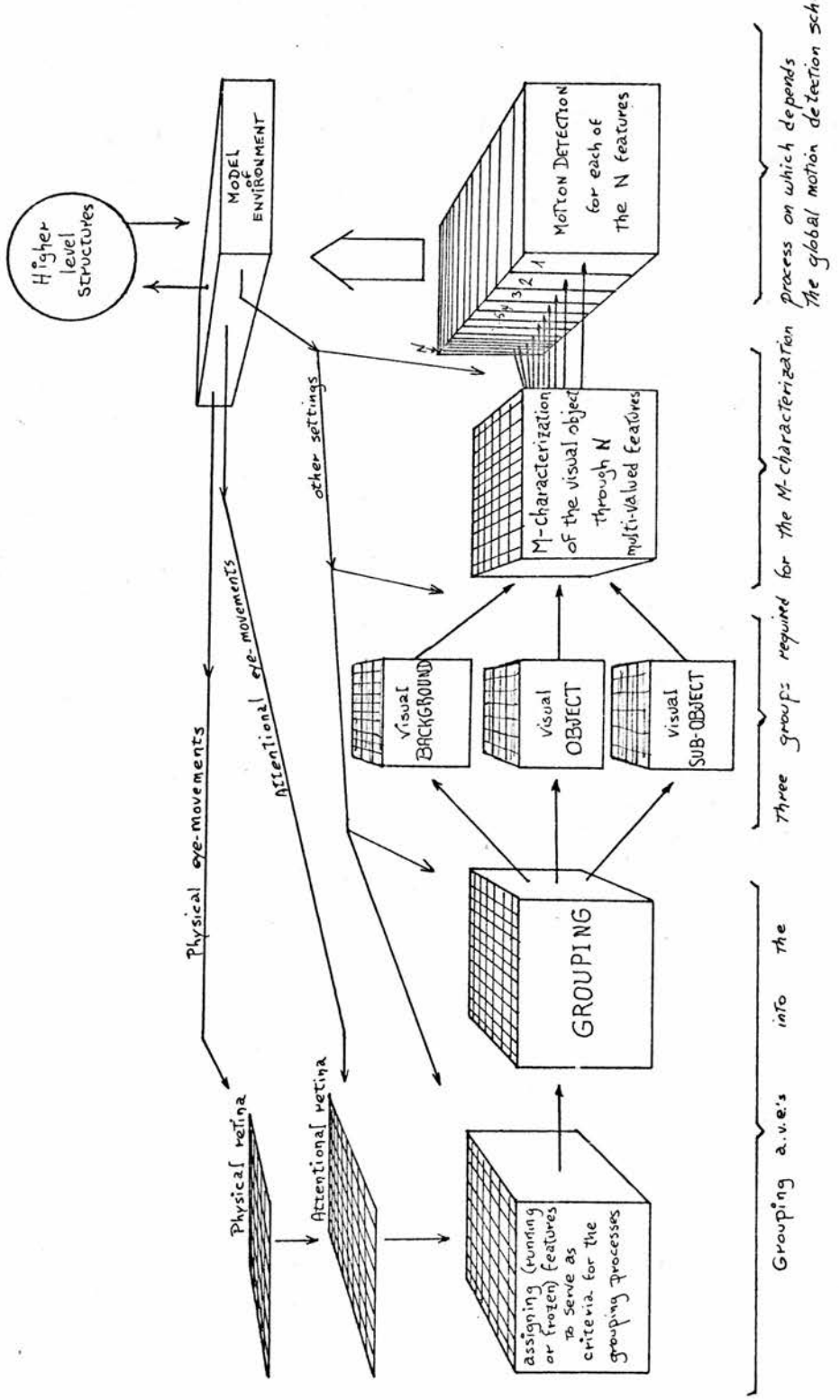
Before trying to focus ideas about the system's general structure by showing how it can be used to tackle a complete scene let us look at a part of this structure which is still unspecified. Up till now, we have ignored totally the problem of deciding how and under which criteria the system should make the groupings which actually yield the three sets of a.v.e.'s needed to achieve motion detection as planned. At this level we come across yet another drawback of the "group first" solution, a drawback not mentioned in the context of the original dilemma presented at the beginning of this chapter because it did not bear on motion detection as such. As the system stands, motion is computed after the three crucial groups of a.v.e.'s have been formed, so it cannot serve as a grouping criterion for forming them. The problem is that we want our system to take some

aspects of motion into consideration when grouping a.v.e.'s into higher level units (cf. discussions of motion field effects in Chapters I and II). The apparently simplest way of achieving just this at the present stage is to equip the system with a local motion detection scheme, operationally totally distinct from the global motion detection scheme which was discussed in the above paragraphs, and meant to provide the grouping system with the desired "running criteria". However this motion detection scheme at (or close to) the level of a.v.e.'s will be subject to most of the drawbacks discussed in the context of the "compute motion first" solution. Fortunately a possible alternative solution was found to this problem of getting running criteria before global motion detection is carried out. It came from realising that motion, or rather "velocity", is not the only possible running feature computable by a visual system and that local change is a sufficiently powerful running feature to provide the system with the running criteria required for grouping purposes. This means for instance that the two groups of a.v.e.'s to be separated in Scene III (figure 4) will actually be separated on the basis of local changes alone, without recourse to motion detection at all. In principle, all groupings required in cases of "motion field effects" should be successfully tackled on the basis of some local change or other. This explains

why we prefer to refer to these cases as "running" instead of "motion" field effects, "motion" restricting the scope of the problem to a too narrow subset of running features, namely velocities. In allowing for "running field effects" on the basis of such a simple criterion as "local change" interest lies in the surprisingly simple computational requirements which this entails (as will be seen in Part II).

So the system now has criteria for separating out the input a.v.e.'s into a background, a visual object, and a visual sub-object. But these criteria are exclusively "running" ones; what about defining some frozen ones? Of course, we do not pretend that running criteria are sufficient in themselves to allow for any desired splitting of a.v.e.'s into the three groups required by our global motion detection system, but since we are primarily concerned with running features and since all parts of the whole motion detection system can be set to work on the basis of some running feature or other we will mainly restrict our concern for <u>frozen</u> grouping criteria (just like our concern for S-characterization outside the scope of Identification in the context of motion detection) to designing the system's data and process structures in a way which makes them as suitable for frozen as for running dimensions.

Figure 10. The system's macro-structure



A general schema of the system is given in Figure 10. To help focus on ideas about how such a system can be made to tackle a scene, we will discuss it in the context of the scene starting with the frame in Figure 11 (presented for instance on a CRT screen). It is important to realise here again that our system is not expected to have any knowledge about what the physical obects portrayed in the scene represent in human terms (i.e. our system cannot know by itself what a lorry, or a plane, or a tree looks like). When we refer to any one of these objects we mean "the set of a.v.e.'s" making it up.

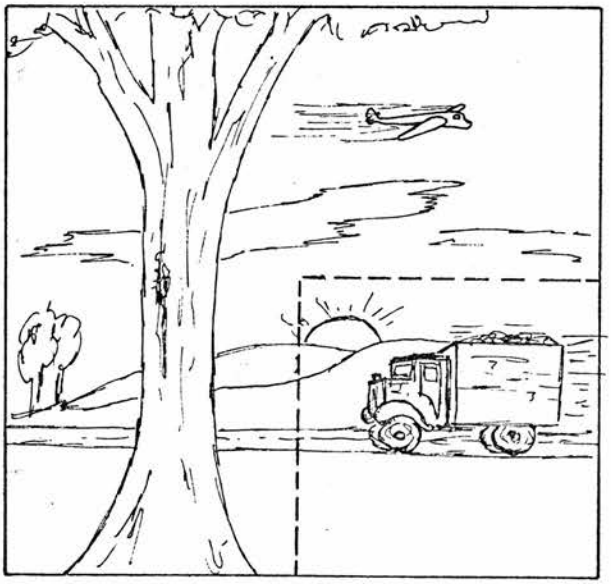


FIGURE 11. Initial frame of lorry/airplane cartoon.

Let us first get the system to transfer all a.v.e.'s detected by the physical retina (on which the whole Scene of Figure 11 is projected) onto the attentional retina, which is the most trivial attentional strategy. From the moment when the attentional retina is loaded with the a.v.e.'s, the system starts trying to circumscribe a visual object within them, using frozen and/or running criteria. However, in this first moment of analysis, since no previous frame was analysed, there is no basis on which to derive interestingly discriminative running features to serve as running criteria. We could therefore concentrate on frozen criteria but, as argued earlier, since we are interested mainly in the running domain, and since the system's ways can be made as clear within this domain as within the frozen domain, let us take in a second frame and allow running features to come into action. Let this second frame be similar to the first one except that the lorry and the plane have both moved forward (in their respective directions). On running grounds, the only possible grouping of a.v.e.'s in the first moment of the analysis was to put all a.v.e.'s in a single box, the visual object's box, thereby leaving an empty set for the background --the way in which the background is always defined is that as soon as the system has decided which a.v.e.'s belong to the visual object, all a.v.e.'s <u>not belonging</u> to this object (and only these)

are automatically assumed to belong to the background. With the second frame, however, due to the changes that occurred in the picture relative to the first frame, the system can create a running basis on which to split the "macro-object" of the first moment of analysis (with the first frame only). In this second moment, the changes undergone by a.v.e.'s making up the plane and the lorry provide a basis on which to put these a.v.e.'s in the "visual object box", while a.v.e.'s making up the rest of the scene (where no changes have been detected) are automatically thrown in the "background box". Moreover, a visual sub-object can be defined, still on running grounds alone, on the basis of the different movements of the a.v.e.'s making up the plane, and those making up the lorry, so that the lorry (or rather the a.v.e.'s making it up) for instance can be thrown in the "visual sub-object box". In this second moment we therefore have all a.v.e.'s not belonging to the plane and lorry sitting in the "background box", all those belonging to the plane and lorry sitting in the "visual object box", and those belonging to the lorry alone in the "visual sub-object box". This is a little more interesting than the classification based on frame 1 alone but it is not quite what we want yet.

The visual object which we have at this point, consists of

all a.v.e.'s belonging either to the plane or the lorry, without any distinction being made between those belonging to one and those belonging to the other, and the system can only analyse the motions of this object as a single entity. This ability to treat any group of a.v.e's as a unit is one of the most important aspects of the system, but to be interesting it has to be accompanied by some capacity to choose which "whole" will be considered as a single object at any given moment. For instance we would like the system to be able to isolate the lorry of our example as single object whose motions get computed. Notice that at moment-2 it was decided to isolate the lorry as visual sub-object in the scene, but the status of "visual sub-object" only allows the a.v.e.'s which bear it to be analysed relative to the visual object as such, which is far from allowing the system to build the desired description of the lorry alone; in fact the visual sub-object is only meant to represent an inherent aspect of the object investigated, not an object itself. So how are we going to provide the system with the ability to work on any desired visual object at any desired level? This will be achieved in three complementary ways;

1. by allowing the system to have its three critical groups of a.v.e.'s of every moment exchange status as moments go by,

2. by allowing the system to select relevant parts of the scene using the combined action of its physical and attentional retinas, and

3. by changing grouping strategies.

Given any scene, different combinations of these three aspects of the system's ability to choose a visual object can yield the same result. For instance, in the case of our scere at moment-2, a change of status is definitely in order if we want to analyse the lorry as such : we want the lorry to go from the sub-object box to the object box. In order to do just that we can either change the grouping strategy (by saying for instance that at moment-3 all a.v.e.'s behaving in a similar way to an a.v.e. chosen from the moment-2 sub-object box will go into the object box), or perform an attentional eye movement which will make the attentional retina bear on the smaller region defined by the a.v.e.'s in the sub-object box (e.g. dotted contours in the scene of Figure 11), or both at the same time. The interest in attentional eye movements is in the fact that they can generally reduce greatly the complexity of the grouping criteria required to yield the desired object. For instance, once the attentional retina is brought to bear on the region defined with the dotted lines already referred to (in figure 11), the simplest

change undergone by a.v.e.'s relative to the retina is sufficient to throw the lorry and only the lorry into the visual object box. The "running" object identification is by no means affected by "broken lines" or "noisy line junctions", or even indeed "occlusion", which have been and still are the nightmare of people working on frozen scene analysis (e.g. when the lorry passes behind the tree drawn in the foreground of Figure 11, a.v.e's moving "on each side of the tree" are still grouped together as belonging to a single object). Once the lorry is considered as the visual object, the visual sub-object box can be filled in with any of its parts, thereby allowing the system to analyse at will the "shape" of the new object. If the system feels that a part of the lorry deserves closer attention, it can change once again the status of a.v.e.'s contained in each of the three boxes, pushing the lorry from the object box into the background box, pushing the desired part (e.g. the door of the lorry) from the sub-object box into the object box, and leaving the sub-object box available for any part of the new object (e.g. the window of the door of the lorry). Conversely, the system can climb up from a very local spot in the scene to a completely global view of it. As the system travels through a hierarchy of visual objects, expectations can be built on the basis of the detected values of the features of these objects and can be used to

direct the computations of the values of those same features in either the sub-object or the background at any given moment. Of course, since scenes are generally made of many very distinct physical objects, the system's analysis does not consist exclusively of this vertical search mostly useful in analysing a complex but single object. For instance, the plane and the lorry were not very usefully described in terms of a single object of which they formed the two main parts. But by isolating the lorry the system can realise (on the basis of some previous knowledge available from higher level structures for instance) that there is no need to fit this object in the context of a larger object, and can limit its hierarchical investigation to the lorry itself. When it is satisfied by its description of this object it is free to move squarely to the plane and start investigating it as a single object. So the attentional retina can be used for moving to a new object (horizontal search) as well as for "digging deeper" into a given object (vertical search). Now as the system searches vertically and horizontally, dealing with one object at a time, all the results of its analysis can be used to build a model of its whole environment. This model is actively used to direct the search itself, in terms of physical as well as attentional movements of the "eye" through the scene and in terms of choices of grouping criteria and expectations

of all kinds facilitating and driving the analysis (referred to as "other settings" in the schema of Figure 10). The model built from the analysis of our "lorry and plane" scene should then, after a reasonable time by human standards, contain vertically and horizontally related objects, possibly identified by higher level structures as being a lorry, or a plane, or a door, or a wheel, etc., specified by their respective motions and their respective "components", some of these components themselves being specified by their motions and/or their components, etc.. These objects potentially stand in this model relative to each other as well as relative to the rest of the scene; and even the rest of the scene, given adequate frozen grounds on which to organise some grouping, could in the model be represented in terms of vertically and horizontally related objects like trees, trunks, branches, leaves, sun, hills, road, sun, clouds, etc.. We do not claim to have brought ways of achieving all desirable groupings, or to have specified modes of higher level intervention, the scope of the solution presented in this chapter being limited to achieving groupings on running grounds and detecting relevant motions. But as far as these latter aspects are concerned, we claim that the solution proposed is adequate for any scene whatsoever, from the simplest case of Figure 2 to the more elaborate case of Figure 1, passing through

intermediate cases such as the above discussed scene involving the lorry and the plane. As a last effort to express the mood of the proposed solution let us now see quickly how Scene III (Figure 4) and the set of Scenes VI and VII (Figure 7) are tackled by the proposed system.

The case of Scene III (Figure 4) is quickly settled. The whole scene gets mapped onto the attentional retina and when the second frame is processed (half of the a.v.e.'s having moved down, and the other half having moved right) the (running) grouping strategies can easily isolate one group of a.v.e.'s as being the visual object; but at this stage the system only knows that there are two groups behaving differently, the actual behaviour of each group being yet unspecified. This specification is achieved in the next step where the group of a.v.e's chosen as visual object is globally M-characterized with successive different positions and "seen" to be moving, for instance, downwards with a specified speed. When the behaviour of this first group of a.v.e.'s has been detected in such a way the system can easily, in the next few moments, analyse the other group's behaviour and find that it is moving rightwards at a specified speed. What has to be realised here is that the global motion of the two groups of a.v.e.'s <u>is not</u> derived from the local velocity of each

individual a.v.e..

Now, the case of Scenes VI and VII (Figure 7) is a rather interesting one. The trouble with these scenes is that they seem to require a level of S-characterization which is sufficient to allow the system to distinguish between the local "shapes" used in the Scene (namely a square, a circle, a triangle, and a cross sign). The lack of powerful S-characterization schemes in our motion detection strategies (based on the fact that since there is only one visual object analysed at a time there is no need for fancy identification strategies) might seem to be a mistake in this case. The fact is that our system can easily do without those specifications of the local shapes used in Scenes VI and VII, and still come up with the desired interpretation (with in fact probably much less computation than the solution through local S-characterization would have required). In the case of Scene VI, the desired interpretation (i.e. the square alone moves from the left to the right hand side of the remaining shapes) is achieved on the basis of the fact that from the first to the second moment only those a.v.e.'s making up the square undergo "some change"; this is sufficient to put them in the object box (all remaining a.v.e.'s being put in the background box), thereby allowing the square to be M-characterized as a whole and

"seen" moving from one side of the "background" to the other. Similarly, in the case of Scene VII, the desired interpretation (i.e. the whole set of shapes being shifted slightly to the right) is achieved on the basis of the fact that from the first to the second moment all a.v.e.'s in the scene undergo a change (with the possible exception of the very few a.v.e.'s which stand in common retinal locations when the different shapes are overlapped through time, but this does not affect the final outcome). This is a sufficient criterion to put them all in the visual object box (leaving the background box with an empty set), thereby allowing all shapes to be M-characterized as a whole which, further analysed motionwise, is "seen" as being shifted to the right.

What has to be stressed here is the fact that quite a lot can be achieved without the help of powerful S-characterization specifying in details all "objects" within the scene, the power of our system resting on the weaker but certainly easier and probably sufficient M-characterization scheme and grouping strategies discussed throughout this chapter. It is within these limitations that our choice of the "group first" solution has to be understood, the grouping being only pushed as far as required to obtain a single but copiously M-characterized object requiring only minimal

S-characterization for Identification purposes. In other words our system allows motion detection to get under way well before the whole visual system has got any idea of what the frozen scene exactly consists of, i.e. what physical objects are represented within it; as a matter of fact not only can our system start analysing motion before the frozen scene is analysed in any detail, but of course it can also provide information for the task of interpreting the scene in its frozen dimension.

The way in which Scenes VI and VII are tackled by our system, i.e. without any recourse to shape identification, has an interesting bearing on Kolers' theory (see pp.96-100). Having acknowledged the fact that human visual motion detection usually makes little use of shape information to decide "what goes where", Kolers "reluctantly" had to soften his position on the basis of experiments using scenes such as VI and VII, thinking that the only way to deal with such scenes was shape-Identification. Our system's way of solving such scenes shows that not only is such shape-Identification totally unnecessary, but there exists a computationally much simpler way of doing it. Moreover it offers a theoretical basis for resolving puzzling inconsistencies (which he acknowledges himself) in Kolers' interpretation of some observations. Having acknowledged that shape is

taken into account to a certain extent, in cases such as Scenes VI and VII, Kolers wonders at the fact that in the case of scenes such as the one represented in Figure 9 (which is not so different from Scenes VI and VII) there is no longer any sign of shape-Identification on the basis of which the system can decide on "what goes where". Figure 12 partly reproduces Figure 9. What happens in the case of this scene is perfectly understandable in the context of our system.

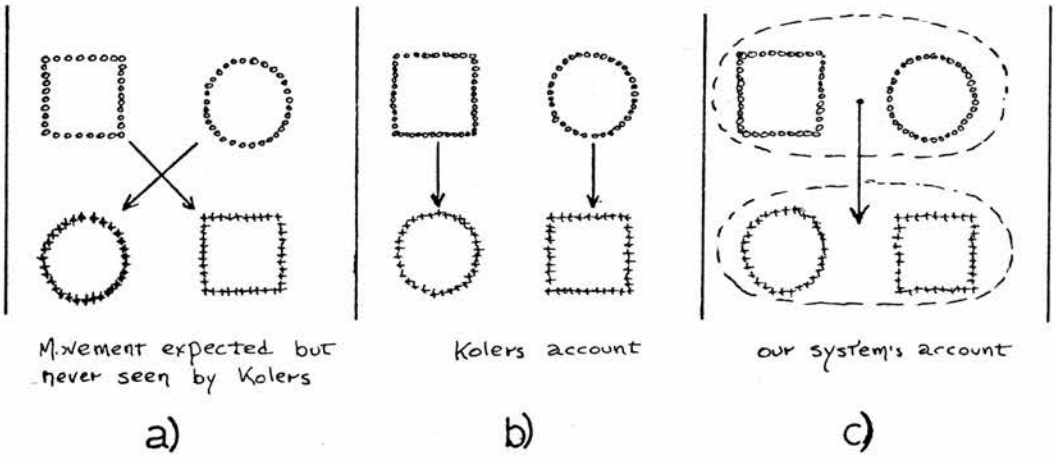


FIGURE 12. Our system's account of the puzzling experimental result obtained by Kolers (1972) in the experiment illustrated in Figure 9 (see p.98).

The way in which our system deals with this scene is that since all a.v.e.'s undergo a change from moment-1 to moment-2, all of them are thrown in the visual object box, they are M-characterized as a whole, and are consequently seen as moving down as a whole. Kolers' "two objects" can

never be made to cross each other in movement because they are welded into a single visual object by our system.

Another interesting aspect of our system as a model of human visual perceptual mechanisms is brought to light in the context of Johansson's theory. As discussed in Chapter II it seems that Johansson's "velocity components" model requires local translatory movements to be detected and then broken into vectorial primitives which allow "common components of movement" to be apprehended as such and act as basis for the grouping. The experiment decribed in Figure 13 brings us back into the context of Johansson's theory.

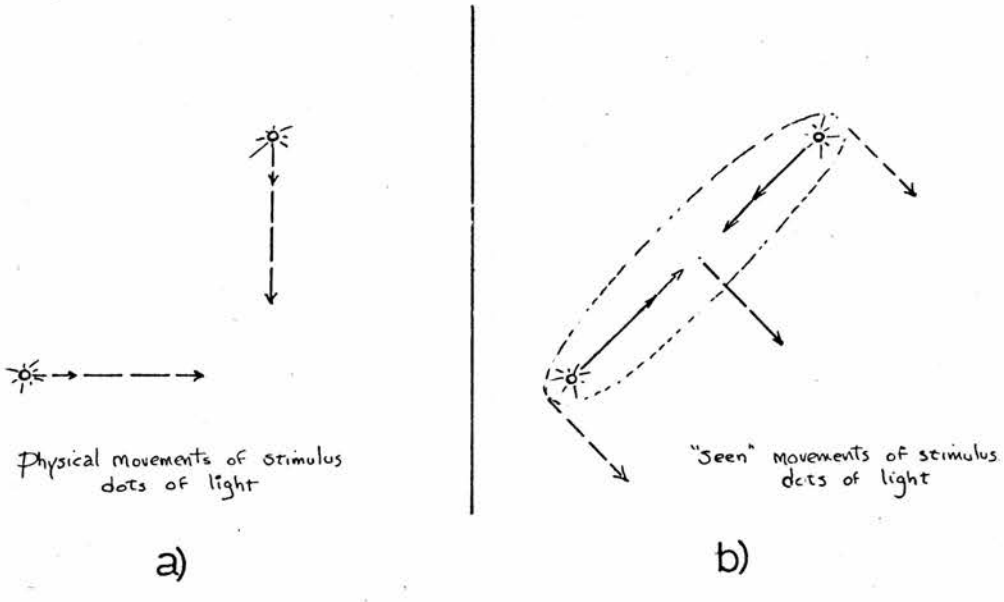


FIGURE 13. Johansson's experiment (1950) to illustrate that the human visual system groups moving objects on the basis of common velocity components.

In this experiment, two spots of light are moving towards each other along orthogonal paths (Fig. 13a); the perceptual "description" of the physical movements is given in Figure 13b. Johansson's argument is that <u>each spot of light</u> is first detected as moving with its physical velocity (relative to the retina of course); these velocities are then broken into components which serve as basis to compare the velocities and extract what they do have in common and what they do not. In the case of Figure 13, it is therefore found that both spots of light have got a velocity component going "south-east", so this common component is "subtracted" from the individual motions and "given" to both spots <u>as a whole</u>, while the incompatible components are left to their respective individual owners. Now in our case, the description given in Figure 13b is also achieved, but since our system does not work with local velocities, the "vectorial analysis" cannot be done as explicitly as in the case of Johansson's model. The way in which our system achieves this result is by doing its global M-characterization. If you M-characterize the two a.v.e.'s corresponding to the two spots of light of Figure 13 with, among other multi-valued features, <u>a global position</u> (which in fact "summarises" each a.v.e.'s local position) the only way in which this global position can change (if of course the global position is determined through a consistent strategy) is

according to the common velocity component of the individual a.v.e.'s. So it turns out that global M-characterization is an implicit way of achieving vectorial analysis of local velocities.

In short, to the first question (how much grouping should be achieved before and how much should be achieved after motion detection?) our answer is clearly that the system should group a.v.e.'s right up to the level of a single visual object (with a background and a sub-object) before motion is computed, and that no more grouping should be done afterwards (within a given processing moment of course). The critical point here is that the "single visual object" need not be the projection of any physical object, which simplifies greatly the grouping task.

To the second question (on which grounds should the groupings be carried out?) the answer is that the system should use both frozen and running criteria to define the visual object which is to be seen in motion. It was furthermore decided that the running criteria to be used should be chosen on the basis of local changes undergone by a.v.e.'s or other low level visual entities.

Finally, to the third question (how should the system go about computing motion as such, i.e. how should the

Identification and M-characterization problems be solved?) the answer is that the system needs not worry about Identification (which is trivially achieved on the basis of the existence of the one and only visual object detected at any moment), but that the system should be provided with means of achieving quite a substantial M-characterization of the object.

Now that our system's macro-structure has been made explicit, we shall turn, in Part II, towards how the computational tasks which it calls for in rather abstract terms can be made to rest on a precise micro-structure or set of effective decision procedures.

PART II

THE SYSTEM'S MICRO-STRUCTURE

or

Where the proposed system's underlying clockwork is revealed

CHAPTER IV

Defining some general purpose primitive concepts [1]

IV.0 Introduction

This chapter defines basic concepts on the basis of which any particular micro-structure underlying the macro-system discussed in Part I can be designed. It is on the basis of these concepts that one such micro-structure will be designed in the next chapter. These primitive concepts will be presented at two different levels of thought (in Sections IV.1 and IV.2); the first level of presentation will be that of intuitive ideas, where the new concepts will be described in plain language and in a rather informal way, and the second level of presentation will be reached by bringing intuitive ideas to a level of precision allowing their embodiment into nerve nets capable of actually performing the required computations.

---

(1) An abridged version of this chapter was presented at the international conference on Artificial Intelligence and the Simulation of Behaviour held at Sussex University in July 1974.

IV.1 Basic intuitive ideas

Here the question is how much of the different grouping and characterizing strategies implicitly or explicitly called for throughout Part I's account of our system's macro-structure can be made explicit in terms of common or general-purpose computational concepts. In other words we want in this chapter to pinpoint explicit processes and/or data structures which cover sufficiently wide non-trivial computational requirements to play a part in the computational "genealogy" of many of the different features derived by the system. By "genealogy" we mean the history of the different relationships between v.e.'s which are taken into account by the system as it proceeds from the level of a.v.e.'s right up to the level of the global visual object and its M-characterizing features (e.g. as the system proceeds from dot-like a.v.e.'s to line segments, from line segments to plane surfaces, from plane surfaces to three-dimensional bodies, etc.). These processes and data structures which belong to the computational "genealogy" of many different features will be called "primitives". In trying to define primitives the idea is to try to account for as many different derivable features as possible through as few primitives as possible. The better one succeeds at this task the more homogeneous the resulting system is, with the ensuing

advantages of relative structural simplicity, speed of processing and compatibility of description between possibly many different actual or hypothetical systems in the same domain.

From our discussion of the system's macro-structure we know that processing has to start from a.v.e.'s; therefore detecting retinal positions and moments of light occurrences (which are the two main characterizing features of a.v.e.'s) are the most primitive processes. As already discussed above, they are dealt with structurally through the sampling strategies. On the basis of the primitive a.v.e.'s, as we saw in Part I, the system can set its grouping strategies to work along two main dimensions, viz. the temporal dimension and the spatial dimension. This leads to the fundamental distinction between running and frozen groupings and their related running and frozen features. Given these two well-defined but different types of features one might well ask on the one hand which feature(s), if any, belong(s) to the genealogy of all <u>frozen</u> features, and on the other hand which one(s) belong(s) to the genealogy of all <u>running</u> features. Since a.v.e.'s are the starting point of all processing and are frozen entities it is obvious that their characterizing features, namely retinal position and moment, are the <u>primitive frozen</u> features we

are looking for. However, even though we know that all derivable running features ultimately set their roots in some primitive frozen features (cf.section I.1), we still have to find a running feature which belongs to the genealogy of all running features.

The most primitive running feature rests on an overlooked primitive frozen feature, namely the "existence" of any detected value of any feature. The reason why this particular feature was overlooked in the discussion about primitive frozen features is due to its trivial nature, inherited from the fact that "existence" is common to all detected values of features within our system and so offers almost no grounds on which to derive in any interesting way what the system is presented with. However, as soon as one starts thinking about what seems to be required to set up grouping schemes, "existence" crops up as the seemingly only possible frozen basis on which a suitable primitive running feature could stand. The frozen feature "existence" is a single-valued feature, i.e. "something" either exists (1) or does not exist (0). What our most primitive running feature will be asked to convey is the "changing" or the "non-changing" in the state of existence (1 or 0) of any value of any feature from one moment to the next. Since this new feature is concerned with pairs of existence states (a pair

consisting of one of the two possible existence states of any given value of a feature for each one of two successive moments) there are only four possibilities, or values, which our new feature has to be able to identify, viz. 0-0 or the "still absent" value, 0-1 or the ON value, 1-0 or the OFF value, and 1-1 or the "still present" value. Our primitive running feature is therefore a multi-valued feature with four possible values, and since these values essentially represent modes of existence through time we will call the feature "transistence". Since we could not find any derivable running feature which could do without transistence in its computational genealogy we decided to consider it as most primitive common running "ancestor" of all derivable running features.

Transistence is a key-concept in the design of our system's micro-structure. Besides playing an essential role in motion detection, and an even more important role in the theory of running groupings, it is the key to the detection of many types of field effects; for example, in Chapter III (cf. p.109) we were talking about transistence values when we were arguing that "local changes" are a sufficient basis for the detection of those apparently relatively complex visual phenomena classically described as "motion field effects".

Transistence can be computed for any value of any feature. This means that values of frozen as well as of running features, of multi-valued as well as of single-valued ones, and values of transistence itself are all perfectly suitable for the purpose of computing transistence. However, it seems that computing transistence on the only value of single-valued features is much less interesting than computing it on the many values of multi-valued features if one is trying for more grouping on the basis of the outcome of transistence computation. Take for instance the single-valued feature "concavity"; once the system has computed transistence on concavity's only value, very little else can be done with the outcome of this computation. Indeed, such values as "concavity ON" or "concavity OFF" or "concavity still there" or "still not there" do not seem to provide a very rich foundation for higher or more global levels of analysis. The point about looking for multi-valued features when one wants to push for further grouping has already been made in Section I.1 (cf. p.21), and it means that if we are to attain higher level running features than transistence we have to concentrate on what can be done on running grounds with the outcome of computing transistence <u>on the different values of multi-valued features</u>. This brings us directly to the next most general type of running feature, and this is <u>motion</u>.

Motion has been loosely described as a running feature resting on the detected values of some multi-valued feature (M-) characterizing some v.e. as time goes by (e.g. translatory motion rests on the changing values of the multi-valued feature "position" characterizing some v.e. at every moment). Let us try to define a little more clearly what this means. As opposed to single-valued features where the existence of a single value confers a global character to its detected transistence value, multi-valued features, because of the fact that they cover many different possible values, only allow transistence values to have a local character. What we mean by "local character" of transistence detection in the context of multi-valued features is that transistence is computed for particular values only, bringing no information about what is happening globally within the whole pool of values belonging to the feature concerned. This of course means that there is room in such cases for some further grouping and more global characterization (for as long as the multi-valued feature concerned possesses individual values which bear some detectable inter-relationships). In other words once you know for instance that concavity has gained or lost existence there is no room left to say anything more global and interesting about the event, but once you know that position has lost a value (OFF) but gained a different one (ON) for some given v.e. there is still

room to ask for more global information like "can these two values of position be related or grouped in some way under a single feature and can their relation be described in any interesting way?" The answer is of course yes and motion is the single more global feature asked for in this case. Motion will indeed be essentially concerned with grouping local values which are detected as being OFF with local values which are detected as being ON within some given multi-valued feature's pool of values. This means that transistence will be used as grouping criterion for motion detection. Now what can be done about characterizing or specifying different groupings of this kind? Characterizing motion of course amounts to velocity detection. This implies the detection of two multi-valued running features, one concerned with a more "qualitative" assessment of the "difference" between the values grouped, that is the direction of motion, and the other with a more "quantitative" assessment of this "difference", that is the speed of motion. In other words, if we consider for instance the multi-valued feature "orientation", with values ranging from 1 to 180, a motion within this feature (i.e. a rotation) could be something like this: at moment-1 orientation 45 turns OFF and orientation 46 turns ON, at moment-2 orientation 46 turns OFF and 47 turns ON, at moment-3 orientation 47 turns OFF and 48 turns ON, and so on. Computing motion in such a case consists in

identifying which value goes OFF and which value goes ON at any moment, and in deriving from them the fact that nothing has globally disappeared or appeared but that something has moved clockwise (i.e. the qualitative relation between OFF and ON values) at a rate of one unit of resolution per unit of time (i.e. the quantitative relation between OFF and ON values).

The computational principles introduced in the above paragraph can be summed up by saying that motion detection implies

A- grouping two v.e.'s characterized by two different values of some given multi-valued feature under the criterion that one of these two characterizing values has an OFF transistence value and the other has an ON transistence value, and

B- characterizing the more global (running) v.e. so obtained through both a qualitative and a quantitative assessment of how different the local v.e.'s characterizing values are.

These computational principles are the general principles to which all motion detection should conform, different types of motions being obtained by applying them to

different critical multi-valued features (i.e. M-characterizing features), e.g. applying them to "position", "orientation", and "speed" will yield respectively "translations", "rotations", and "accelerations or decelerations". In Chapter V we will consider which particular features should be chosen to cover the range of relevant physical motions in the environment and how each one of them could be treated in accordance with our general motion detection principles. The main point here is that whatever critical multi-valued feature happens to be chosen for the grouping, this grouping and the ensuing characterization will have to conform to the general computational principles reported above. This means that however "unrelated" different types of motion may seem to be on the basis of their respective critical features, they will all have in common the general motion detection principles discussed in this section.

There is more to be said about what all types of motion detection have in common; the scope of what remains to be said actually reaches beyond strict motion detection and applies to all groupings, frozen as well as running. What we are concerned with now is data structures, whose design will rest on the following points. Firstly the assessment of relations between different values of a single

multi-valued feature, such as velocity detection, is the essence of visual grouping strategies in general, and secondly defining processes for relating values of multi-valued features could be greatly facilitated by working within data structures where the different values of the multi-valued features concerned are already set in a way which implicitly accounts for the relations to be made explicit.

Concerning the first point, relating values of multi-valued features in order to reach more global information about some set of v.e.'s is by no means the exclusive prerogative of motion detection, even though it is one of its main concerns. The grouping process on which motion detection is based gets its exclusive character from the fact that it groups pairs of values of (multi-valued) features consisting of one OFF value and one ON value. Putting aside these grouping criteria, and their consequence that the v.e.'s grouped belong to two different moments, the analysis of relations between the values grouped is a strategy available in both the frozen and the running domains. This common charactristic of all groupings should be exploited as much as possible in designing any particular data structure.

Now concerning the second point, about the advantage of

having data structures where useful relations between values of features are "available" from the storage lay-out itself, a most obvious example is our system's two-dimensional input array of retinal receptor cells from which, for instance, topological adjacency of two detected local positions of light occurrences can actually be derived (or detected) by pairing structurally adjacent signals specifying detected positions of light. The idea is to provide such structures for as many multi-valued features as can be derived by the system, and one way of doing this is by creating within the system multi-dimensional arrays where each dimension is devoted to the organized storage of the different values of some given multi-valued feature.

Taking our first point into consideration what we propose is to design as many common "organized multi-dimensional arrays" as possible for storing multi-valued features' pools of values. A great advantage of having such data structures is that for those multi-valued features which lend themselves to running as well as to frozen analysis the system offers a common workshop for deriving the required new frozen and running features. But whatever the advantages of our views about data structures may be, the main point to be made here is that these views rest on general-purpose concepts affecting the computation of many

different types of features, and that the main concept is that of having as many multi-valued features as possible represented through multi-dimensional arrays where each dimension consists of an organised mapping of all the detectable values of one multi-valued feature. Such multi-dimensional arrays will be used extensively in designing our particular system's micro-structure and will be referred to as "piles".

Let us now conclude by reviewing quickly the primitive concepts introduced in this section. Firstly, a.v.e.'s together with their two main characterizing features (viz. retinal local position and moment) were acknowledged as "most primitive common ancestor" of all derivable features. Because of its frozen nature this "most primitive common ancestor" of all derivable features was also acknowledged as "most primitive common frozen ancestor", that is as most primitive frozen entity in the genealogy of all derivable frozen features. A feature called "transistence" was found to be the most primitive common running ancestor of all derivable running features, taking root in the trivial primitive frozen feature "existence" and extending it through the temporal dimension into a multi-valued feature specifying the four possible pairs of existence states which any value of any feature can bear through two successive moments: 0-0 (the

"still not existing" pair), 0-1 (the ON pair), 1-0 (the OFF pair), and 1-1 (the "still existing" pair). Going further into the running domain the most primitive common motion ancestor of all motion features was found to be the grouping of pairs of v.e.'s characterized by different values of the same multi-valued feature under the criterion that one of the values has got an OFF transistence value while the other one has got an ON transistence value. Furthermore, all such groupings were said to imply a characterization on the basis of both a qualitative and a quantitative assessment of the difference between the OFF and the ON values. The generality of this strategy of deriving more global features from assessing differences between different values of a single multi-valued feature characterizing more local v.e.'s was then acknowledged. Finally general comments regarding desirable data structures in which to carry out these assessment processes led to the concept of a "pile", or multi-dimensional array where each dimension consists of a mapping of all values of one given multi-valued feature and where the possibly relevant relations between the values of each given multi-valued feature are available in the structure of the mapping itself.

IV.2 Nerve net embodiment of basic intuitive ideas.

This section will deal with effective decision procedures and actual data structures which express in an unambiguous way the computational concepts introduced intuitively in the last section. We want these more precise versions of our primitives to be simple enough to allow for high speeds of processing (by human standards), and we want them to be precise enough to be directly implementable (explicitly in hardware or implicitly through computer simulation) wherever required in an actual working system. This new description of our primitives will be spelt out in terms of nerve nets.

The main basic concepts, or primitives, introduced in the last section are those of

1-a.v.e.'s as most primitive entities and their characterizing features (position and moment) as most primitive detectable features,

2-transistence as most primitive running feature,

3-motion and velocities as next most primitive running features, and

4-piles as general-purpose data and process structures for storing and grouping values of multi-valued features.

The case of a.v.e.'s and their characterizing features is easily settled since they have already been said (in Section I.1) to be derived structurally on the basis of our system's basic sampling strategies. In terms of nerve nets this simply means that there has to exist a set of as many cells as there are different retinal positions of light to be detected, each cell being specific to one and only one position and vice-versa, and that each one of these cells is either activated or not within a specific period of time. Each such cell represents, at any given moment, a particular a.v.e. with its particular position and moment and existence state. It is from the information available from this set of cells that the system has to start its grouping and characterizing task.

With transistence the task is a little more demanding. From what we have said in the last section, detecting the transistence of some value of some feature involves a procedure which takes as input the existence state (either 1 or 0) of the value at some given moment and, by pairing this existence state with the one detected the moment before, produces as output one of the four possible values of transistence. An effective procedure which does just this is expressed by the nerve net shown in Figure 14, where signals travel the distance between any two "nodes" in one moment and where any node sends at any given moment

a signal in every one of its output lines if it has received at the preceding moment a number of incoming signals greater than or equal to its threshold. The "number of incoming signals" is given by the algebraic sum of activating (+) and inhibiting (−) signals reaching the node at any particular moment.

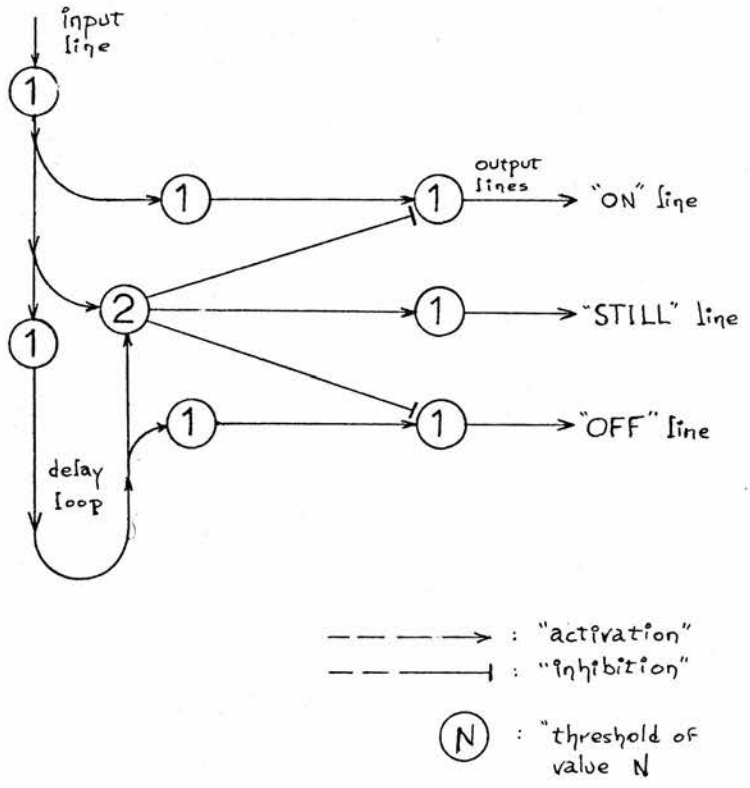


FIGURE 14. Transistence detection unit (TDU).

Figure 15 shows a precise situation where the network actually computes values of transistence: at moment 1 the given feature's value is absent, i.e. its existence state is 0, and there is no signal already running in the network; at moment 2 the given feature's value is present

(i.e. it turns ON); at moment 3 the value is present again (i.e. it remains "still"), and at moment 4 the value is absent (i.e. it turns OFF). The expected outcomes of the net's computations are obtained respectively at moment 4 (ON verdict), moment 5 ("still" verdict), and moment 6 (OFF verdict).

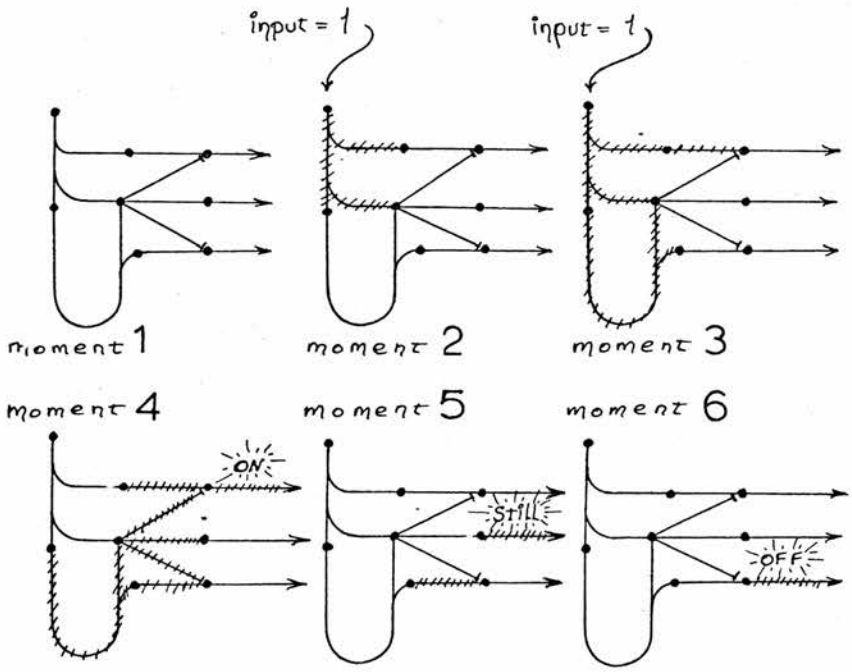


FIGURE 15. TDU at work.

We can see from this example that each possible matching (of the input at one moment with the input at the moment before) is represented by a specific outcome in the

network: 0-1 (the ON value) is specified by a signal in the upper output line, 1-1(the "still present" value) by a signal in the middle output line, 1-0 (the OFF value) by a signal in the lower output line, and 0-0 (the "still absent" value) by "none of these signals". Obviously one and only one of these four possibilities is realised at any moment. The computation is achieved by using a delay loop to keep in the network the input received the moment before (confirming the running nature of the grouping), and by a simple combination of activating and inhibiting signals controlled by thresholds at particular junction points to carry out the matching process and generate the specific output.

This effective procedure, through its nerve net expression, will be called for whenever and wherever transistence needs to be computed and it will be referred to as a "Transistence Detection Unit" (TDU).

Before moving on to discuss the nerve net embodiment of motion and velocity computing the following points should be made concerning the TDU.

First we want to stress the fact that since the TDU is designed to compute transistence on particular values of features, if the system is to be kept informed of what is

happening to every detectable value of any given multi-valued feature we have the choice between providing the system with one single TDU to be treated as "sub-routine" which is called to compute transistence on each value as the system exhaustively goes from one to the next, or providing the system with an individual TDU for each possible value, thereby making parallel processing possible. The computational simplicity, not to say triviality, of the TDU allowed us to choose the much more satisfying parallel setup, which means that if the feature considered has N possible different values there will exist N different TDU's, each one being specifically linked to one particular value.

Secondly we want to emphasise the general-purpose character of the TDU, that is the fact that transistence can be computed for any detected value of any detected feature, including of ccurse transistence values themselves as detected values of a detected feature.

And thirdly we want to make it clear that we do not propose the TDU as an anatomical unit that should be found in actual nervous systems. The network which we are proposing is exclusively intended to be a <u>conceptual</u> tool to tackle the theoretical problem of running groupings. In other words any resemblance with any existing natural

anatomical network is a pure coincidence.

Now what about our third main primitive concept, that of motion itself and of its two characterizing features, direction and speed? We saw in the last section that motion was a running grouping involving different values of some multi-valued feature, the criteria for such a grouping having to be found in the transistence values of these different values of the multi-valued feature. The critical transistence values were identified as being the OFF and the ON values, and the characterization of the outcome of the grouping carried out according to this criterion was required to cover both the qualitative and the quantitative differences between the values whose transistence served as grouping criterion. Since transistence is required as grouping criterion we clearly want to use TDU's as a starting point. Also, since all values of the multi-valued feature concerned are potential elements for the grouping, the system ought to have every value of every multi-valued feature for which motion is to be detected (i.e. every M-characterizing feature) provided with its specific TDU, the output of these TDU's providing at every moment the required grouping criteria. Given a particular M-characterizing feature with its pool of values and its associated pool of TDU's, the only thing that remains to be done in order to allow the computation

of any motion "undergone" by this feature is to provide the system with a related network where any OFF value of the feature can be coupled with an eventual ON value of the same feature in a way which makes qualitative and quantitative differences between the two values apparent. Such a network can be thought of as being a set of "channels" mapping out all the possible interesting journeys through the feature's pool of values and associated TDU's, where each channel would stand for a possible "direction" of motion and where the length of each channel would be available from the number of TDU's having access to it. The idea is that whenever a value of the feature turns OFF, the OFF signal triggered by the TDU associated with this value is let loose in the motion detection network, being sent through all channels in search of an ON signal. As the OFF signal progresses through each channel it counts the number of TDU's which have access to the channel but which are "silent", thereby computing the distance being travelled through the feature's pool of values. Whenever an ON signal is met in any particular channel a "motion" signal is generated, the particular channel in which the meeting occured stands for the direction of the motion (i.e. the qualitative difference between the OFF and the ON value), and the distance travelled in this particular channel prior to the meeting stands for the speed of the motion (i.e. the

quantitative difference between the OFF and the ON value). Notice that only OFF signals are "travelling" through the network, the ON signals simply being made to "cross" channels at points which are specific to the respective values which they characterize. For this reason the above described channel network will be called the "travelling OFF network", and will be globally labelled as a Velocity Detection Unit (VDU).

Let us "see" in more visual terms how this VDU works. Figure 16 shows what the network would look like for a single direction and from a single TDU's point of view (i.e. only one TDU's OFF signal can travel along the line in search of an ON).

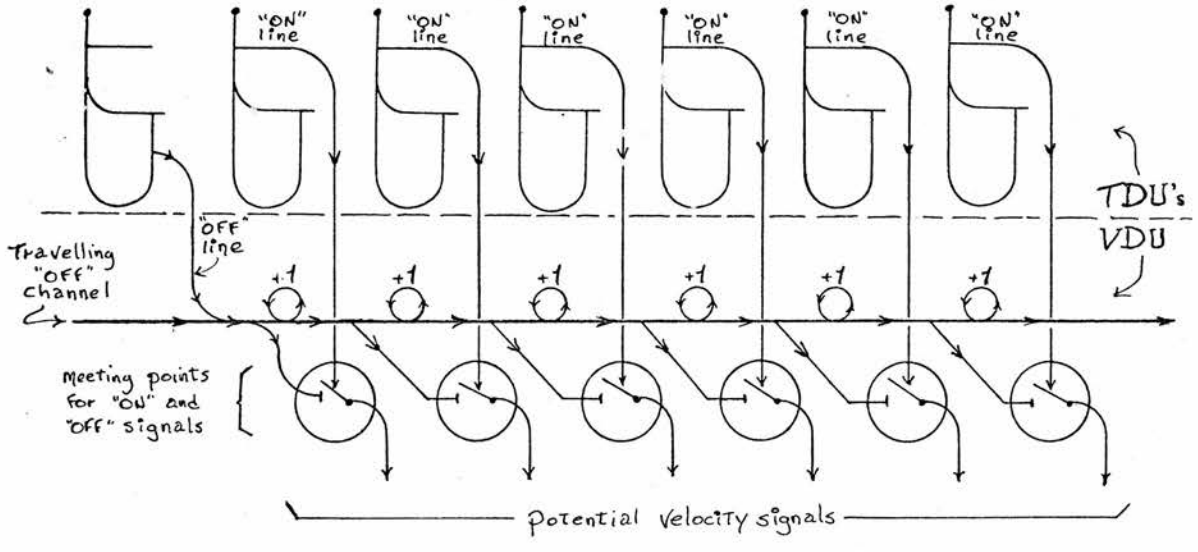


FIGURE 16. Velocity detection unit (VDU).

Figure 17 shows an example of how the VDU would work in the case of an OFF signal computed by the first TDU and an ON signal computed by the fifth TDU.

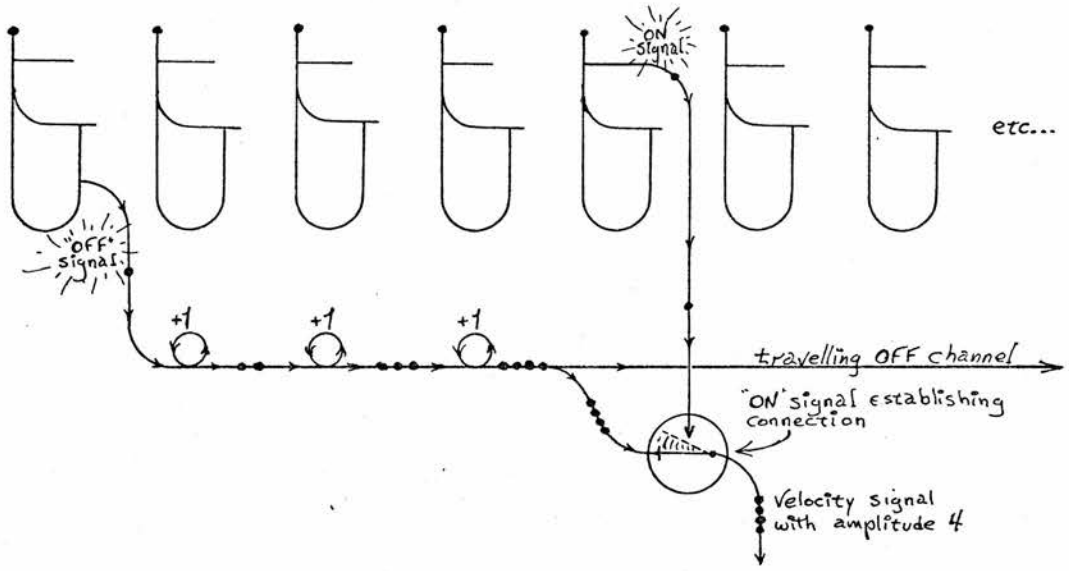


FIGURE 17. VDU at work.

Let us choose the particular feature "orientation" and see how it can be fitted with a VDU, i.e. with means of deriving "rotations". To start with, we assume that all values of the feature "orientation" are equipped with their respective TDU's; the number of values pooled under the feature "orientation" will depend on the resolution which this feature is required to have i.e. if the chosen resolution is 1 degree then there will be 180 different values. Whenever an OFF signal is triggered by a TDU attached to some value, since values of orientation can only change in two main ways (viz. clockwise or

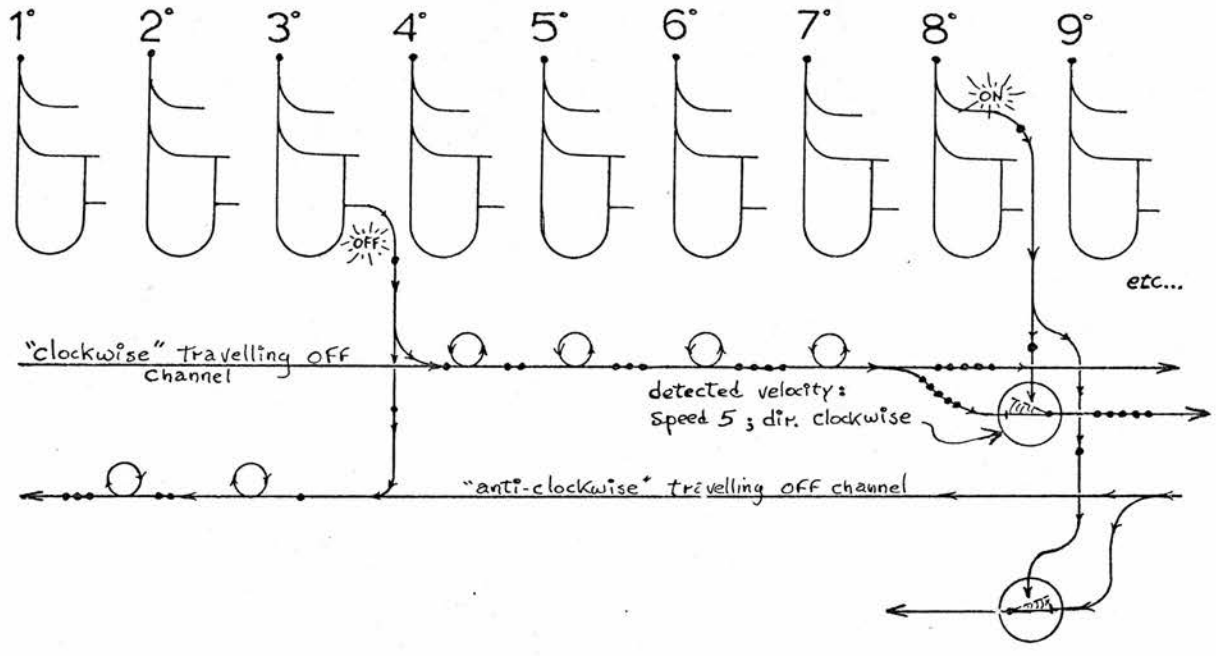


FIGURE 18. "Rotation" VDU.

anti-clockwise), this signal will be sent along two travelling OFF channels in search of an ON signal. The OFF signal will count on the way the number of "silent" TDU's which it passes, and since each TDU passed by stands for one unit of resolution of <u>orientation</u> the "distance" computed will be rotational. If the OFF signal meets an ON signal somewhere along the clockwise "travelling OFF" channel then a clockwise motion will be reported by the VDU, and if the meeting occurs in the other channel an anti-clockwise motion will be reported. If no ON signal is met then the OFF signal gains a global status over the

whole pool of values (the same happens for an ON signal not met by an OFF signal). Figure 18 shows how such a system would work for two TDU's only, an OFF signal being detected at 3 and an ON signal at 8 .

Now there are two very important remarks to be made about this general velocity detection scheme.

The first is that given any M-characterizing feature on which the system is to compute motion, if at any moment there is more than one OFF signal and/or more than one ON signal detected in the feature's pool of values the system is at a loss in deciding which OFF is to be paired with which ON. This problem is at the very heart of the motion detection issue, and finds its highest level extension in the Identification problem discussed in section I.2. It is mainly with this problem in mind that we decided to opt for the "group first and compute motion afterwards" solution in the design of the system's macro-structure. This solution was taken to the point of asking for a single visual object to be defined with sufficiently numerous and descriptive M-characterizing features on which to compute motion. This duly characterized single object is our guarantee that no M-characterizing feature on which motion is to be computed will allow for more than one OFF and one ON signal at any given moment in the VDU

underlying its pool of values. This should not be interpreted as meaning that our motion detection scheme cannot cope with more than one moving physical object at a time; the point is that due to the complexity of the control systems which would be required to drive efficiently a search involving many OFF and many ON signals it seems to be a good idea to try to do without this extra complexity and limit VDU's to one OFF and one ON signal at a time. Of course many motions can be computed in parallel: we will allow for at least as many VDU's as there are different M-characterizing features specifying our single visual global object.

The second remark is that the travelling OFF network, or VDU, required to make our motion detection scheme efficient asks for highly structured pools of values (and corresponding TDU's) on to which to be grafted. If the particular features' values (and TDU's) through which the OFF signal has to travel are not set in such a way as to be structurally positioned to facilitate the setting of pathways for the travelling OFF network, direction and speed detection might become too complex. If for instance all the values of orientation shown in Figure 18 were set in random succession instead of the ordered way in which they are portrayed, setting the travelling OFF network would not be the simple matter which it is in the case

shown. This brings us directly to our fourth main primitive: piles.

From what was previously said about piles and what has just been said about the required structure of pools of values through which OFF signals are to look for ON signals it seems that "piles" is what our VDU's should be anchored in. The piles within which the VDU's are to stand should therefore be multi-dimensional structures where each dimension consists of one M-characterizing feature's pool of values (and related TDU's) whose layout expresses in a linear way the possible interesting directions and distances between detected OFF and ON values. It should be noted, however, that it is only because OFF and ON signals are used as grouping criteria that we can talk of VDU's within such piles, the use of other criteria transforming the same piles into process and data structures for deriving other features (one of the main advantages of piles). The main point about nerve net piles at this stage is that all storage and processing concerned with different values of multi-valued features will be "dimensionalized" structurally through rows and layers of nerve networks where spatial adjacency of nerve cells will be made to bear a "featural adjacency" of some kind and where linear journeys through these rows and layers should allow all required levels of global

representation to be reached. Actual nerve net structures fulfilling these general requirements will be described in details in the next chapter.

Let us briefly sum up the main points arising from this section's discussion.

Firstly, we made the basic input material to our system, viz. a.v.e.'s, available through a set of nerve cells triggering a signal or not at any given moment depending on whether or not the local retinal position which each one of the cells specifically represents receives light. It is from those signals (standing for a.v.e.'s) that our system should derive new visual entities and features through running and frozen groupings.

Secondly, we expressed the computation of the most primitive running feature, viz. transistence, in terms of a precise nerve network, the TDU, whose computational simplicity is sufficient to allow its actual grafting onto any value of any feature whose transistence value at any moment might be required.

Thirdly, the computation of motion itself was also expressed in terms of a precise network, the VDU, where motion is detected on the basis of OFF and ON outputs from

TDU's specifically grafted onto the different values of a given multi-valued feature, and where the motion is characterized through a velocity (a direction and a speed) worked out by specific journeys through the structured network.

Fourthly, it was decided that the structured networks required to make VDU's viable for velocity detection should be implemented in terms of multi-dimensional arrays of nerve networks, or "piles", providing linear arrangements of cells representing in a highly organized way the different values of detected multi-valued features.

CHAPTER V

Designing the system's micro-structure

V.0 Introduction

In this chapter we will try to use the primitive concepts introduced in chapter IV to articulate a working visual motion detection system along the general lines of the macro-structure proposed in Part I. We will therefore concentrate on designing precise mechanisms to group a.v.e.'s into a single global visual object and to compute its possible motions.

The presentation of the system's micro-structure will involve two main steps. First, in Section V.1, we will deal with the parts of the system concerned with grouping a.v.e.'s into a single visual object (and its two "satellite" visual entities, viz. a background and a sub-object), and then, in Section V.2, we will deal with those parts of the system concerned with M-characterizing the visual object and computing actual motions. In both sections, as we did in Chapter IV, we will start by giving an intuitive account of the problems and solutions before presenting precise nerve net embodiments.

V.1 From a.v.e.'s to the single visual object

V.1.1 Basic intuitive ideas

From our previous discussion we can say that there are two main "macro-decisions" with which the processes to be described here should conform. The first is the decision that all a.v.e.'s selected to go on to our system's attentional retina should ultimately be grouped into only three sets: two disconnected ones, the "background" set and the "object" set, and a third one consisting of a sub-set of the "object" set, the "sub-object" set. The second decision is that the grouping criteria to be used in defining these three sets should include running ones, and that these running criteria should consist of computationally simple features such as "local change". Our problem in this chapter is to transfer the appropriate a.v.e.'s from the physical to the attentional retina, and from there to define grouping schemes involving at some stage "local changes" as criteria and ultimately leading to the three final frozen visual entities, namely the background, the object, and the sub-object.

Our first concern will be for the part of the system which lies between the input material, available at every moment from the physical retina, and the selected part of this

input material, lying on the attentional retina. The very low level at which this attentional retina is introduced in the system, i.e. prior to all main steps in the processing, leaves the system with very little scope for actually selecting what should go on to it. It is however felt that allowing the system to select any one region of its physical retina as object of analysis is sufficient to start with, and that since placing the attentional retina next to the physical one does allow such a selection we should postpone discussing the advantages of placing it deeper into the system until the need arises. So the relationship between the physical and attentional retinas will be kept rather straightforward; at every moment the system will decide, on the basis of whatever information it possesses at that time, which region of the physical retina deserves attention, and all a.v.e.'s falling within this region on the physical retina will be transferred onto the attentional one. The only problem with this scheme is that as the system will be devoting all its computational power to a particular chosen part of the physical retina anything happening on the remaining part of the retina will be lost. However there is no need for our system to disregard completely what is happening on its physical retina outside its field of attention. On the contrary, it seems that it would be useful for the system to be notified whenever and wherever "something"

happens on its physical retina outside its field of attention. In such cases the system could transfer its attention immediately and have a closer look at the event. So all we want is a warning that "something" is happening, and an indication of where it is happening, the system being free to follow up the warning or not depending on its attentional priorities. Transistence computed on all positions of the physical retina can provide such a warning at minimal computational costs. So we propose that transistence should be computed for every a.v.e. on the physical retina before the chosen sub-set of these a.v.e.'s on the attentional retina becomes the sole object of analysis.

Once the attentional retina has been loaded with the chosen set of a.v.e.'s the system's task becomes one of generating and characterising new v.e.'s by grouping these chosen a.v.e.'s. Our task is one of deciding exactly how this grouping and characterising should be carried out right up to the level of the three v.e.'s required for motion detection. Our main concern is to decide how and at which level of the grouping scheme running criteria should be introduced. The deriving of frozen grouping criteria will be confined to those frozen features which are necessary to the deriving of the running grouping criteria which we are interested in, although we will make

every effort to provide all the required facilities for eventually including frozen feature computation in the same framework.

The first point requiring clarification is the exact nature of the running criteria which we wish the system to derive. It was decided in Part I that there was no need to go as far as motion to obtain a sufficiently powerful running basis on which to do some interesting grouping, that in fact "local change" could be a perfectly suitable and computationally much simpler basis. But what does "local change" mean? What we had in mind when using this expression in Part I was "transistence", and our claim is that transistence is powerful enough to provide the system with running grounds for coping with running field effects or other groupings requiring running criteria.

Having chosen transistence values as running grouping criteria in deriving our three final frozen v.e.'s (i.e. background, object, and sub-object) the next step is to decide at which level(s) of the grouping process and on which features transistence should be computed in order to yield appropriate grouping criteria. Doing it at the level of the attentional retina's a.v.e.'s seems rather dangerous in that it could restrict us to changes which only have a meaning at a very low level, as in the

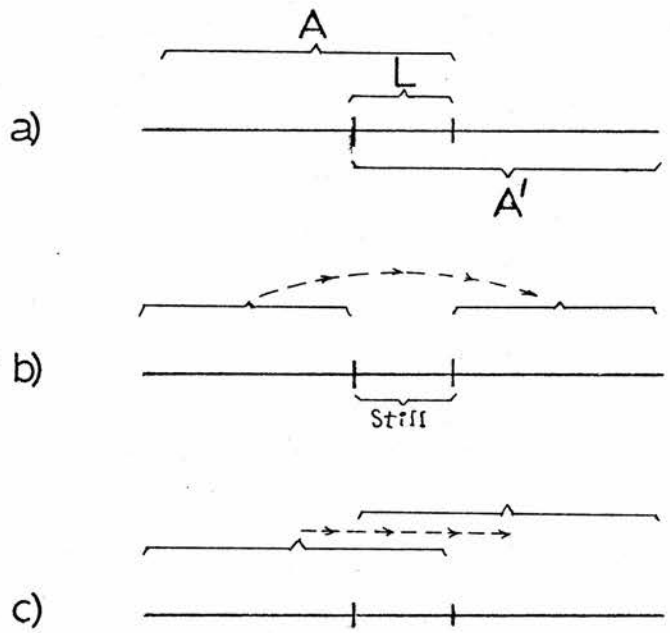


FIGURE 19. The case of the straight line moving along its own axis.

following case. Referring to Figure 19a we have a situation where line A represents the set of the a.v.e.'s at moment-1, and line A' represents the set of the a.v.e.'s at moment-2, that is we have a situation where a line segment is moving along its own axis from the left to the right. At moment-2 all the local positions (i.e. a.v.e.'s) corresponding to the line "A minus L" are detected as having an OFF transistence value, all those corresponding to the line "A' minus L " are detected as having an ON value, and all those corresponding to the

line L are detected as having a STILL value. The grouping of a.v.e.'s according to these transistence values can only lead to the interpretation shown in Figure 19b, which can only mean for a motion detection system that a line (A) gets split into two smaller lines, one of which moves to a new position, jumping over the other one which remains still. Surely we do not want such a result; what we want our system to see is a single line A moving as a whole over the distance "A minus L", as shown in Figure 19c.

One way of obtaining such an interpretation would be to give all the a.v.e.'s (making up the whole line at every moment) the status of a single unit with only one global position, and to compute transistence on this single unit's position instead of computing it on each of the unit's local elements (a.v.e.'s). In other words, it seems a better idea to compute transistence on the positions of line segments than on the position of a.v.e.'s which make up the line segments. We are of course talking here of "continuous" line segments, i.e. line segments consisting of adjacent a.v.e.'s. It is indeed mainly because of this adjacency that we run into trouble with the situation shown in Figure 19, the adjacency of the elements (a.v.e.'s) of the moving object having made it impossible to detect some local changes by

causing many retinal positions to be continuously occupied even though the line was moving on. Choosing to group only those a.v.e.'s which are adjacent on the attentional retina means that situations like the one represented in Figure 20a (where the set of dots marked A,B,C,D,E and F is presented at moment-1 and the set of dots marked A',B',C',D',E', and F' is presented at moment-2) will still be interpreted by our system as shown in Figure 20b (which is the "discrete" analogy of the case shown in Figure 19b) but this time, for some reason, the interpretation does not seem unacceptable at all, it even seems quite plausible.

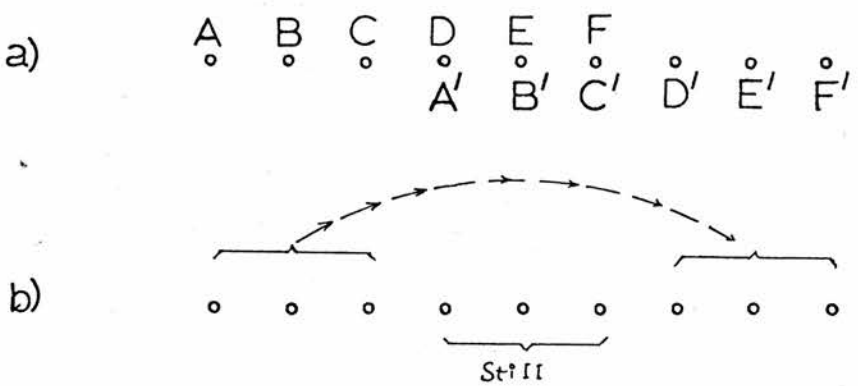


FIGURE 20. The case of the disconnected aligned dots.

Having discarded a.v.e.'s as interesting objects for transistence computing in our search for running grouping

criteria, we are now faced with the task of designing some process for grouping a.v.e.'s into higher level v.e.'s, namely line segments, on the frozen grounds of adjacency. This incursion into the frozen domain should provide us with visual entities and features on which to compute the transistence values required for our grouping purposes and should also be a step forward in the purely frozen domain.

The problem is to find ways to group a.v.e.'s under the criterion of retinal adjacency. Adjacency can be derived from the different positions (critically) characterizing a.v.e.'s at any moment, but the problem is that a.v.e.'s can get to be adjacent in many ways, so that different characterizing features might have to be spelled out in order to allow for an adequate characterization of the resulting group, or v.e.. We will limit the types of adjacencies leading to such v.e.'s, or line segments, by considering as "legal" line segments only those sets of a.v.e.'s where the spatial relationship between any <u>two</u> adjacent a.v.e.'s in a set is the same as or the converse of the spatial relationship between any other <u>two</u> adjacent a.v.e.'s in the same set, and where all a.v.e.'s in a set belong to a single "chain" of adjacent elements. This restriction on the types of adjacencies that can lead to line segments means that the system will only consider unbroken <u>straight</u> line segments. So, only three

characterizing features will be required to cover the range of all possible instances of detected line segments, viz. position, orientation, and size. All three features are of course multi-valued (frozen) features, and therefore lend themselves to motion detection. Although they are excluded by our restrictions on types of adjacencies, curved line segments are by no means brushed aside; we are only waiting until we are in full control of straight line segments before changing the "two's" to "three's" in our stated restriction on what could be a line segment and before adding curvature to position, orientation, and size in our set of line segment characterizing features.

What we want now is a process by which all a.v.e.'s on the attentional retina are split into as many groups as there are potential line segments in their layout, and by which every such group is given one position value, one orientation value, and one size value. Of course we want this process to be parallel, efficient and quick.

The whole issue about detecting straight line segments revolves around deriving their orientation. The "spatial relationships" which were said to be the basis for deciding if two a.v.e.'s are adjacent in a "legal way" have to be available somehow, and their diversity will be

a direct function of the potential power of the system at detecting different orientations of line segments. The fact is that in order to satisfy our criterion of grouping all those a.v.e.'s which are adjacent in a common way, we have to go through, at least implicitly, detecting the "way" in which they are adjacent in order to be able to say that this "way" is common to all a.v.e.'s grouped, and "ways" cannot represent anything else than orientations. Once a line segment has been detected, and a particular orientation given to it, position and size can be derived almost trivially by working out any consistent strategy to raise a local element's position to a global status (e.g. the "middle" one) and by evaluating the number of elements (a.v.e.'s) which the actual group, or line segment, consists of.

So how are we going to tackle orientation? The problem of orientation detection mainly lies in the fact that given any set of similar retinal cells it seems impossible to find a retinal geometry where any one cell has got more than six adjacent neighbours (i.e. there are only six different possible relationships between any two adjacent cells on the retina). Of course we want our system to be able to detect many more different orientations than this. The problem is that with any retinal geometry just about any orientation which is not explicitly accounted for by

the natural adjacency of the cells will correspond to an "irregular staircase" or "zigzag" type of arrangement on the retinal cell structure, and such arrangements do not provide legal adjacencies in the context of straight line segment detection. What we therefore have to do before applying the criteria for detecting line segments is to transform the "staircase" arrangements which correspond to strict linear orientations into strict linear arrangements, and this can be done by providing the system with structures for mapping retinal cells in a way which creates "pseudo retinal arrangements" where the adjacency criteria for straight line segments will be applicable. In other words we are proposing some kind of template matching scheme. We believe that template matching is always the best solution whenever the input diversity to be handled is relatively small: it is efficient, quick, and lends itself perfectly to parallel processing. Since a resolution of one degree of orientation would only require 180 templates, a template matching scheme for detecting line segment orientations can easily be adopted.

Since we will want to express the results of the line segment detection scheme in piles, and preferably in one pile where the different dimensions represent the line segments' characterizing features, there is no reason why the pile itself should not be used as template matching

space where the dimension representing the different values of orientation provides the set of required templates to actually detect any line segment's value of orientation. Since furthermore curved line segments can be considered as being sequences of partly overlapping straight linear segments in adjacent orientations the straight line segments detection pile would be the ideal structure within which to derive curvature.

So now that we have an idea of how line segments can be detected and characterised, and of where the results of these processes can be found, we can come back to the problem of deriving running criteria for reaching the level of our three ultimate frozen v.e.'s. Line segments' characterizing features are position, orientation, and size. Each one of these features has a pool of values whose transistence can be computed and can serve as criterion for further groupings. The question now is can we gain anything from computing transistence on the values of such features. This question can be readily answered by placing ourselves in the context of a cartoon such as the one presented in Figure 1 (p.9) showing a lorry physically moving relative to a physically still background. In such a case a simple computation of the transistence value of the position of every line segment in the picture would be sufficient to split line segments

into three definite groups: the "still position" line segments (i.e. the background), the "OFF position" line segments (i.e. the lorry at the previous moment) and the "ON position" line segments (i.e. the lorry in the current moment). Of course the system has no idea at this point that there is a lorry there; it only knows that there is a group of line segments which "behaves" as a whole. This is fine, but such transistence values computed on "position"'s pool of values are only a help in those cases where part of the scene is moving while the other one is still relative to the system's retina. In those cases where two groups of line segments move relative to the retina as well as relative to each other, our transistence scheme seems to be at a loss. And such situations are extremely common. One instance is that of our lorry moving in one direction relative to the retina, and the background being "pulled" in another direction relative to the retina. Clearly the line segments making up the background and those making up the lorry behave as wholes and as independent wholes, but our scheme cannot detect that. Another example is the random dots situation used to make the case for movement field-effects (cf. Fig.4, p.39). It might seem that in such cases motion is needed to detect the field-effects. In fact there is no need to introduce such computational complexity in the scheme. The solution is that transistence can be computed

on relative positions of v.e.'s as easily as it can be computed on their absolute (or retinal) positions. The only thing that requires to be done is to choose one line segment's position as reference, to plot all detected line segments' positions relative to it at every moment, and to compute transistence on the plotted positions instead of the original ones. Using such a scheme, the line segments making up the lorry can be separated from those making up the background, and the dots going right (in Fig.4) can be separated from those going down, and all this is achieved with utmost computational simplicity. So the system will compute transistence on both the absolute and relative positions of line segments, and it will use the results of this computation to seperate out the object's line segments from the background's. We could move on to discuss how the same scheme can be applied to line segments' orientation and size, but absolute and relative positions seem to give us sufficient power for the time being. We can obviously come back to these unexploited possibilities if need arises.

We now have means of providing specific groups of a.v.e's (expressed in the more compact form of line segments) standing for visual object and background, but we have yet to provide criteria for specifying which of the line segments making up the object can stand for sub-object.

This can be done rather easily. Once the system has decided which line segments are to form the object, a scheme similar to the one used to separate out background line segments and object line segments can be used to find out which object line segments should also be sub-object line segments. The only difference in this case is that since sub-object line segments are only interesting in their relation to the whole object, there is no need to compute transistence on the object line segments' absolute positions. So as soon as the system has decided which line segments will form the object, transistence will be computed on their relative positions, and the line segments making up the sub-object will be chosen on the basis of the transistence values obtained. An example showing the use of these facilities to create the three frozen v.e.'s is that of a lorry moving relative to a still background but at the same time raising its rear part to unload something. In absolute terms all line segments making up the whole lorry undergo some change while those making up the background do not, so the lorry as object can be separated out from the background; and in relative terms all line segments making up the rear part of the lorry undergo some change, so the rear part of the lorry can be isolated as sub-object.

Summing up the proposed computational steps from the

physical retina's set of a.v.e's at any moment to the three frozen v.e.'s which are to serve as basis for computing motion we find

1-the computation of transistence on the physical retina's positions of a.v.e.'s to provide the system with a warning whenever something happens on the physical retina outside the field of attention,

2-the choice of this field of attention in the form of selecting a precise region of the physical retina from which all a.v.e.'s are mapped on to the attentional retina,

3-the grouping of the attentional retina's a.v.e.'s into line segments characterized by a position, an orientation, and a size,

4-the computation of transistence on the line segments' absolute and relative positions, and the grouping of line segments into an object and a background on the basis of the transistence values obtained, and

5-the computation of transistence on the relative positions of the object's line segments, and the grouping of line segments which are chosen to belong to the sub-object on the basis of the transistence values obtained.

Before proceeding, the following important comments must be made to avoid possible misunderstandings. First of all

let us be clear that on the sole basis of transistence values the system cannot possibly know which detected group of line segments represent the physical background and which represent the physical object any more than it can know what physical entities each group stands for. The important point is that the system does group together line segments which "behave" together. The task of deciding which group should be considered as visual object and of deciding which physical object this group stands for belongs to higher levels of processing. Secondly, it should be stressed that it is only to limit the problem that we concentrated on running features as grouping criteria to define the three global v.e.'s. We realise that meaningful groupings will in most cases require combinations of frozen and running features. However, even though we have been mainly concerned with providing the system with running grouping criteria the line segment detection pile which was used to do so could also be used to derive just about any desired frozen criterion based on line segment position, orientation, or size (absolute or relative). The "twisted pile" could easily be used for instance to derive the different line segment intersections on which Guzman's body identification scheme (1968) is based.

V.1.2 Nerve net embodiments

To start with we need a two-dimensional array of nerve cells acting as physical retina. Following the steps proposed in the last section, our next task is to provide this physical retina's two-dimensional array of receptive cells with nerve nets responsible for issuing warnings of any happening on the physical retina outside the system's main field of attention. This is readily achieved by providing each cell of the physical retina with a TDU. So right next to the physical retina's two-dimensional array of cells the system will have a corresponding two-dimensional array of TDU's, computing each position's transistence at every moment. More importantly, the physical retina's a.v.e.'s will also be given access to a two-dimensional array of cells having exactly the same dimensions as the physical retina but onto which only selected regions of the physical retina can be mapped at any moment, and this array is the attentional retina. Straightforward inhibitory processes in the attentional retina can be used by the higher level control centres to regulate the moment by moment selective mapping process between physical and attentional retinas.

This leads to the first really demanding task, that of grouping the attentional retina's a.v.e.'s into line

segments and characterizing each one of these line segments with a particular position, a particular orientation, and a particular size. As decided in the last section, this whole task should be carried out in the context of a pile within which the result of the processing is stored. The line segment detection and storage pile could be designed as follows.

Consider a pile of arrays on top of which the retinal array is sitting, and let these arrays be square arrays of similar square cells (1). Now let the first array under the attentional retina sit in the same orientation as the retina but let the second one be rotated clockwise (looking down the pile) about its centre by an amount X (degrees); rotate the third array in the same way but by an amount 2X, rotate the fourth array by an amount 3X, the fifth one by an amount 4X, etc...etc...until the rotation reaches 180 degrees (see Figure 21a). When this is done you have in front of you a "twisted pile" of two-dimensional arrays or layers where each such array represents one particular orientation (each array being a

(1) To lend itself to the scheme which we are about to describe we believe that the geometry of the retinal and other arrays can be any geometry of a compact two-dimensional array of geometrically similar elements; square arrays and square cells are only chosen for the ease with which they can be described and simulated.

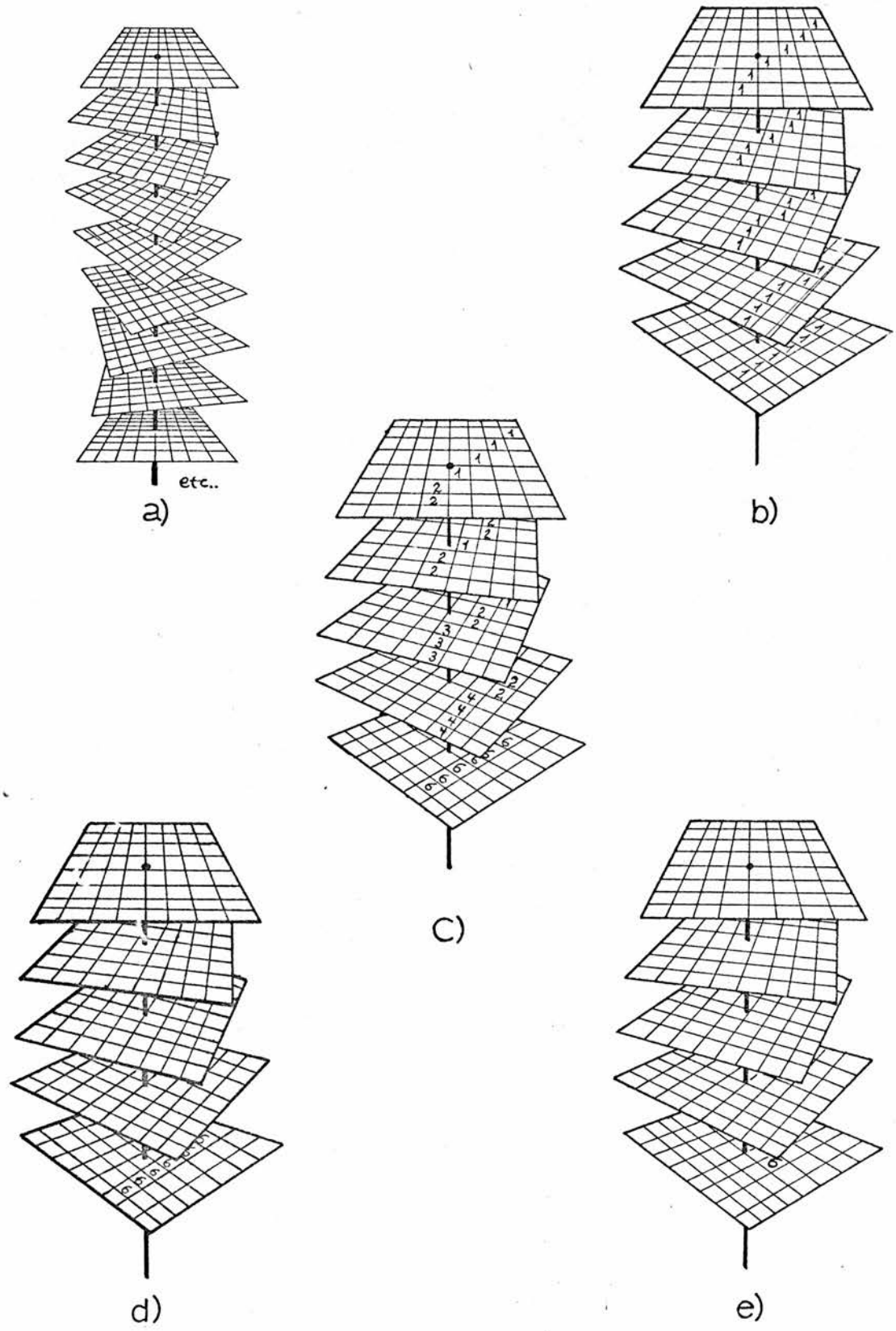


FIGURE 21. The line segment detection pile

"template" for the orientation which it is sitting in). The value which is given to X in building such a pile obviously determines the orientational resolution of the system.

Now you send a "wire" of nerve cells from each primitive retinal position straight down through the whole pile, establishing a connection with each "position" it meets as it goes in a straight vertical line through the successive superposed layers (or arrays); these wires represent the desired connections between the retina (i.e. the attentional retina's a.v.e.'s) and the orientation templates. The way in which this set-up will work is that when detected a.v.e.'s on the physical retina (i.e. those a.v.e.'s whose existence state for the current moment is 1) are plotted in their corresponding positions on the attentional retina, positional signals available from these cells will be sent (vertically) through the pile along their respective wires and a "1" will be put in each cell (or array position) which they happen to be connected to on the way to the bottom of the pile (see Figure 21b). The result is that if the stimulus is a straight line in the orientation A then the array which is sitting in the closest orientation to A will have received the greatest amount of 1's in adjacent cells (or positions) of a single column (i.e. a column in the plane of the layer). In

order to find this "critical" array we can do the following. First of all we add, in every column of every array, the adjacent 1's and we replace each "1" by the sum of adjacent 1's in the set it belongs to; for instance if in a given array in the pile there are six adjacent 1's in one column, then each 1 is replaced by a 6 (see Figure 21c). This allows us to use local positions with "global" weights. We use every value so obtained as local basis for "vertical" inhibition, the idea being to have the larger value (or weight) in any given set of vertically linked positions erasing the smaller ones so that in the end we are left only with the value which is sitting in the array representing the orientation of the line (see Figure 21d). This is done in parallel through all the vertical wires in the pile, or sets of vertically linked positions, so that for every line segment detected a set of values is left on the array representing its orientation. Each set of values, representing each line segment through as many values as the line segment covers positions on the layer, can then be brought to a single position, the middle one of the set for instance, the single cell corresponding to each such position in the pile representing a specific line segment whose orientation is specified by the particular layer in which it lies, whose position is specified by where it stands in the particular layer, and whose size is specified by the

value or weight which it contains (see Figure 21e). This gives only a general idea about how our system's template matching pile for line segments should be wired up. Let us now have a closer look at the vertical inhibition scheme which it calls for.

In the above process, vertical inhibition is meant to destroy the content of all those "occupied cells" in the pile which do not belong to a longest line. Since the critical comparison of lengths of lines is only done vertically by a straight down or up projection of each occupied cell's "weight" many longest lines of different lengths can and will coexist in the pile in different positions in the horizontal dimension whenever the stimulus consists of more than one line segment (as in the overwhelming majority of cases). Concerning the vertical inhibition itself, a first idea is to have each cell in the pile connected to its two immediate vertical neighbours: one above, one below. Given one straight line segment as stimulus, since the length of the projection of this line in single columns of the layers of the pile should decrease steadily as one goes away from the array in the critical orientation, inhibiting only the closest neighbours above and below should be quite sufficient to get rid of everything except the value in the critical array. However, when we tried to simulate our ideas about

the pile on a digital computer we realised that we could very rarely get a perfectly smooth shortening of lines in single columns as we were going further and further away from the critical array. What happened is that at some point in the pile two successive layers would detect lines of equal length in single overlapping columns, and this would create a situation where our inhibition scheme could not get rid of one of them. One solution was to extend the inhibition power of each cell to the whole pile instead of restricting it to the two immediate neighbours. We tried out the idea by making each value inhibit all those values which were on its path as it went straight up and down through the pile. This turned out to be too much: values which should have been kept because they belonged to other genuine line segments were erased in the process. The solution which finally turned out to be satisfactory involved a vertical inhibition decreasing with the distance away from the layer of origin, i.e. initial inhibition weights are decreased as a function of the number of layers by which they are separated from their original layer. The best such decreasing function was found to be non-linear and to follow approximately the cotangent's decrease in value with an increasing angle. In this way the expected results were obtained.

It might be worth noting that the inhibition scheme makes

certain aspects of the structure of the pile look rather inappropriate. Since orientation is a "cyclic" feature (i.e. there is no "first" orientation or "last" orientation), a pile with a top and a bottom is a rather unsatisfactory representation space for it. Consider a pile which twists to 180 degrees, the immediate "downwards" (or clockwise) neighbour of the bottom layer of the pile is the top layer of the pile, its second "downwards" neighbour being the second layer of the pile, while the immediate "upwards" (or anti-clockwise) neighbour of the top layer of the pile is the bottom layer of the pile, its second "upwards" neighbour being the penultimate layer of the pile, etc...etc... In other words the inhibition circuits through the pile would be much more adequately described as ring-like circuits in a ring-like pile (i.e. a pile whose top and bottom are brought together) than as vertical circuits in a vertical pile. So our twisted pile should be thought of as a twisted ring, but we will carry on referring to it as the twisted line segment detection pile.

Although the pile is now equipped to carry out a successful vertical inhibition of all values which are derived from the initial mapping of the attentional retina's a.v.e.'s in each layer of the pile, there are still some problems to be solved before the line segment

detection pile can be considered operational. These problems were also brought to our attention as we tried to simulate the pile on a digital computer and they have to do with retinal resolution and the precision of the mapping of primitive retinal positions into the positions (or cells) of each layer of the pile. Mapping problems in fact brought about two of the most interesting results of this part of the design of the system's microstructure. We realised first that the nature of the pile was such that no position on the retina exactly overlapped a precise position in each array of the pile. Therefore, we had to define a strategy for deciding which position in each array should be connected to each retinal position. We kept obtaining very poor results until we decided to think of retinal positions in terms of fields consisting of many single receptors. The idea of the field is to assign to each position in every layer of the pile a set of retinal receptors (or local positions) which, by reaching a certain percentage of stimulation, sets to 1 the existence state of the particular single position which they are mapped into. The point here is that the retinal receptors can be grouped into fields in many different ways depending on which layer they are being mapped into. A consequence of this new set-up is that there will have to be more receptors (or local positions) on the retina than cells (or positions) in each array of

the pile if we want to keep the same resolution as before. The threshold percentage of cells to be excited within any given field at any moment in order to have the system consider the field's position as being excited was set to 75% in the computer simulation.

The introduction of fields improved the mapping considerably but many lines were still not getting through. The problem was that the relatively high threshold introduced to decide if a field is excited or not (75% of the receptors having to be stimulated) made the system ignore those lines which did not fall more or less exactly on complete fields; for instance when half of the width of a line falls in one column of an array and the other half falls in an adjacent column, this line will not be detected because it only covers 50% of the width of either column. Lowering the threshold did not improve the situation because it introduced all sorts of unwanted detections. We were then left with the possibility of improving the resolution by increasing the number of fields through new combinations of the already existing retinal receptors. However, we found a mechanical way of achieving exactly the same result without imposing any extra load on the computation: tremor, or constant shaking of the retina. The main drawback of this mechanical solution is that it can be rather time consuming because

of its serial nature, but if we can afford the time loss it seems to be the best solution. From experimentation with it on the simulated pile, it proved to be very satisfactory. The tremor was introduced in the following way. Realising that we only needed a displacement of the retina (relative to the stimulus) by an amount of (at the most) three quarters of a receptor field in order to push any line-stimulus into a "slot" (or template) of some array in the pile, we decided that tremor should not exceed sweeps of three quarters of a field in any direction (in fact half a receptive field's amplitude is sufficient). Since it is also clear that the tremor has to allow lines of all orientations to fall in their respective "slots" we had to ensure the multi-directionality of the sweep. We therefore decided that one tremor cycle would only be completed when the retina had swept over an approximately circular region three quarters of a receptive field in diameter. This means that before allowing the vertical inhibition process to be initiated in the pile the system has to wait for at least one tremor cycle to be completed. This provides us with a genuine basis for deciding on the absolute value of the system's basic sampling moment, a decision which was deferred until we could find sensible processing requirements on which to base it. The important point about tremor in the context of the sampling moment is that

if the chosen sampling moment is shorter than the tremor cycle period "false" transistence values will be detected by TDU's as the tremor displaces the stimulus relative to the retina. It seems essential to compute transistence on the result of the summation process which ends with the completion of a tremor cycle. This means that a stimulus (physically) flashing at any rate above the tremor rate will be seen by our system as a continuously lit object. So in our system the threhold for perceived fusion of a flicking light will be determined by the temporal magnitude of the tremor cycle. Since we have already chosen the so-called "gap-sampling" strategy our system cannot accept information at every tremor cycle, so we will accept retinal stimulation for one tremor cycle out of three (i.e. the system will take one and skip two).

Before moving on to discuss the details of how the line segments obtained can be grouped into a visual obect, a sub-object, and a background on the basis of transistence, we will conclude our discussion of the line segments detection pile by emphasising three very interesting characteristics of this structure.

Firstly the highly parallel nature of the proposed processes and their great simplicity permit a high speed of processing. Secondly, and much less obviously, the

line segments detection pile lends itself very well to the computation of curved line segments. The criterion to accept a given straight line in a given layer of the pile is that it should consist of the longest set of adjacent "occupied cells" in one single column of the layer when compared with its projection through the other layers of the pile. If the stimulus is a straight line segment and if its orientation is not exactly half way between the orientation of two successive layers in the pile, there will be a longest set of adjacent "occupied cells" in some layer of the pile; but if the stimulus is a curved line segment (let us assume constant curvature for the moment), there will be a sucession of layers with a longest set (i.e. with equal longest sets) of "occupied cells" projecting into each other, and every set will survive the vertical inhibition process. This is sufficient to specify a curve, but this is also sufficient to detect

1- the actual curvature of the line segment (since the length of each detected straight line in the pile of straight lines specifying the curve will vary with the curvature of the stimulus-line),

2- the length of the curved line segment (since the number of layers with "equal longer sets of occupied cells" will vary with this length), and

3- the orientation of the curve (i.e. the orientation of the tangent to the line segment's middle point).

Thirdly, it might be worth stressing that the twisted pile can be used as a general-purpose tool for running as well as frozen searches for positions, orientations, and sizes, and that it can be used for deriving and representing virtual as well as actual line segments, "skeleton" as well as "contour" line segments, etc...

Now that all line segments lying on the attentional retina are available with their respective positions, orientations and sizes, we can start thinking about grouping them into the three sets required for M-characterization and motion detection. It was argued in the last section that transistence computed on the values of the line segments' characterizing features could provide sufficiently powerful grouping criteria for line segments which "behave" together in the running domain. Let us see how the twisted pile can allow transistence to be computed on line segments' characterizing features. Since every detected line segment is represented in the three-dimensional pile by a particular value in a particular cell, <u>any change</u> in the value of any cell means either a change of position, or a change of orientation, or a change of size. So it is possible, by simply linking a TDU to every cell in the pile, to distinguish between on the one hand line segments whose values of position <u>and</u> orientation <u>and</u> size remain the same from moment to moment

and on the other hand line segments whose values of position or orientation or size change from moment to moment (leaving the actual feature(s) having undergone the change in value completely undefined). Such a scheme is easily implemented but lacks power because of its non-specificity to changes in values of particular line segment features. A more specific setup could be obtained by linking to every cell in the pile a TDU which takes as input at every moment a "1" if the cell contains any value but 0, and a "0" otherwise. Such a TDU, which we will call a "position-orientation" TDU, will not detect any change in size of the line segment represented by the cell to which it is linked, but it will react to either a change of position or change of orientation. An even more specific setup could be made to detect changes of position only. Position-specific TDU's can be obtained by making position-orientation TDU's take as input a "1" if the value of the particular cell to which they belong or the value of any cell directly above or below this cell throughout the pile is a non-zero value, and a "0" otherwise. Such TDU's would react neither to changes of size (i.e. to changes of non-zero values of cells) nor to changes of orientation (i.e. changes of layers of line-segment representing cells), but would immediately detect any change of position of a line segment. Combining such a position specific scheme with a

position-orientation specific scheme would allow the system to identify changes of orientation only. These few examples of different transistence detection schemes should be sufficient to show that the twisted pile does offer quite a rich context for computing transistence values.

For the time being we will limit TDU setups to position-orientation ones. This type of TDU setup is one of the simplest and is quite sufficient to provide running criteria on which to base the detection of some interesting field effects.

First of all let us create a new twisted pile, totally devoted to transistence computing and storing. This pile will be provided with one TDU per cell, each TDU taking as input at every moment a 1 or a 0 depending on whether the particular cell to which it belongs contains any non-zero value (1) or not (0). The results of each TDU's computation at every moment will be stored in a <u>fourth dimension</u> of cells: each cell in the pile, besides being provided with a TDU, will be provided with three cells (forming a fourth dimension within the pile) standing respectively for an OFF result of the associated TDU's computation, or an ON result of this same computation, or a STILL THERE result (the STILL NOT THERE possibility

being obtained by default of the other three) (see Figure 22). So at every moment the transistence value of the position-orientation values of any detectable line segment will be represented by a 1 in the appropriate cell if it is an ON, an OFF, or a STILL THERE value, and by a 0 in all three cells if it is a STILL NOT THERE value. What we are doing in creating such a four-dimensional pile is simply creating embedded three-dimensional twisted piles where the "top" pile is the standard line segment pile, where the next one is the "ON" pile (containing all ON line segments), the next one the "OFF"pile (containing all OFF line segments), and the last one the "STILL THERE" pile (containing all STILL THERE segments).

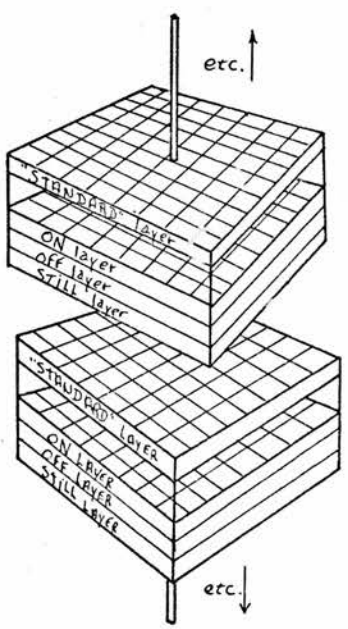


FIGURE 22. How each layer of a twisted pile devoted to transistence computing is equipped with three associated layers specifically used to store the ON, OFF, and STILL THERE transistence values computed for each primary layer cell's content value at every moment.

Now it was decided in the last section that the grouping of the line segments into an object and a background should be done on the basis of either absolute (i.e. relative to the retina) or relative "changes" in the line segments' positions. We therefore propose to use two four-dimensional twisted piles for transistence detection and storage: one for absolute position-orientation transistence, and one for relative position-orientation transistence. Absolute transistence computing is quite straightforward: line segments are taken directly from the line segment detection pile and mapped into their respective cells in the absolute transistence pile where TDU's compute their position-orientation transistence and store the result in the appropriate (fourth dimension) cell. On the other hand relative transistence computing, although it is by no means complex, requires a little adjustment of the line segments' positions before having transistence computed. First of all line segments are transferred directly from the line segment detection pile to the relative transistence detection pile, each line segment being mapped into its specific cell. Then a reference line segment is chosen, and this is the closest line segment to the centre of the retina, after which all line segments in the pile are shifted together by an amount such that the reference line segment then lies in the centre of the retina. Once this shifting procedure

has been carried out the line segments occupy the relative positions on which we wish transistence to be computed, so TDU's come into action and store their results in the usual way. Now when the next moment comes, since we do not want a <u>different</u> reference element to be chosen, we have to look for the <u>closest</u> line segment <u>to the spot where our reference line segment was the moment before</u> (rather than to the centre of the retina), so that in fact the line segment shifting amount becomes cumulative through successive moments, until of course the reference element has moved far enough to justify the start of a new sequence through the choice of the line segment which is actually closest to the centre of the retina for the current moment.

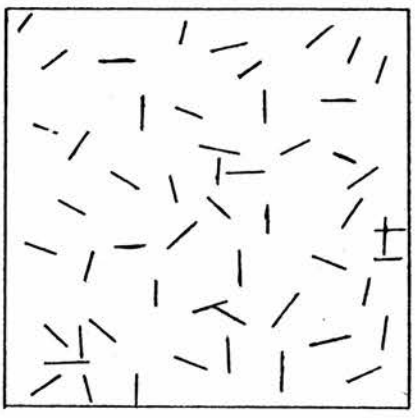


FIGURE 23. Random line segments pattern.

In order to illustrate quickly the type of results which can be obtained with these two types of transistence detection piles let us consider a random elements pattern such as the one shown in Figure 23. In a first case let us imagine that half of the line segments in the display, chosen randomly, move while the other half remain still. It will be easy for our absolute transistence detection pile to put all those line segments which "appear" in the ON sub-pile, all those which "disappear" in the OFF sub-pile, and all those which "do not move" in the STILL sub-pile. If we consider a second case where a randomly chosen half of the line segments move in one direction while the other half move in another direction, the absolute transistence detection scheme will only be able to differentiate between OFF positions and ON positions which is not too much of a help in distinguishing between the two moving groups. But what will happen in the relative transistence detection pile is that one of the line segments will be chosen as a reference, and whatever the moving group to which it belongs happens to be all line segments belonging to this group, because of the shifting procedure, will be detected as being STILL while all line segments belonging to the other group will go in the OFF and ON categories. If these schemes work for random element patterns there is no reason why they should not work for organised patterns.

Now that we have provided our system with means of obtaining running criteria to help decide which line segments should go into the object "box" and the background "box", we ought to provide it with some structure within which to reach just such a decision. This structure will be a twisted pile, but this time its fourth dimension will be much more extensive than in the transistence detection piles. This four-dimensional pile will consist of a "top" three-dimensional pile which will contain the usual line segments' specific cells but "under" which will lie as many other piles as there can be useful or relevant different values of different features on which to base the splitting of line segments into an object and a background. Six such associated piles will of course be the absolute ON, OFF, and STILL piles together with the relative ON, OFF, and STILL ones. Others can be thought of being for instance the line segment intersection (or junction) type pile or any other frozen attribute of line segments. It is by travelling through this pile containing all the useful grouping criteria that the system will generate two different groups of line segments, each group being transferred to a new different pile. Coming out of the object-background differentiation pile we therefore find two new piles: the background pile, encompassing all the line segments which supposedly make up the background, and the object pile,

encompassing all the line segments which supposedly make up the object. It is by travelling through these piles that the system will have to derive features to characterize the global entity which each one contains, but before moving on to this we still have to create and fill in the sub-object pile. This is an easy task which only involves repeating processes which have already been used for creating and filling in the object and background piles. It is even simpler since it was argued in the last section that only relative transistence detection was required to provide the desired running grouping criteria for deciding which line segments are to belong to the sub-object at the given moment. So all we need to do once we have filled in the object box is to transfer all "object" line segments to a relative transistence detection pile where they will be "labelled" with their relative position-orientation value before being taken to the object/sub-object differentiation pile from where a precise sub-set of them will be finally transferred into the sub-object pile.

An example might help to distinguish the different aspects of our system's ability to group together elements of a scene which behave together in the running domain. We will use again Figure 23 as starting point, but this time the random elements of the pattern will be divided into

three (randomly defined) groups one of which remains completely still while the other two are set into translatory motion in such a way that all line segments belonging to each one of the two groups move with a common velocity but where this velocity is different for each group. In such a case our system can use absolute position-orientation transistence to separate out the "moving" from the "non-moving", putting all the still line segments for instance in the background pile and all the changing line segments in the object pile, and it can then separate out the "moving in one way" from the "moving in another way" by simply relying on the relative position transistence scheme allowing the detection of a running sub-object within the object itself.

In conclusion, Figure 24 gives an overall view of the system at this point in its development. Notice in the schema the different levels where instructions from higher level centres are needed or allowed in: a first level is that of the choice of which part of the physical retina is to go on to the attentional retina (taking for granted the possibility for higher level centres to choose which part of the environment the physical retina will be exposed to); a second one is the choice of the reference line segment for the computation of the relative position transistence in the context of the object/background

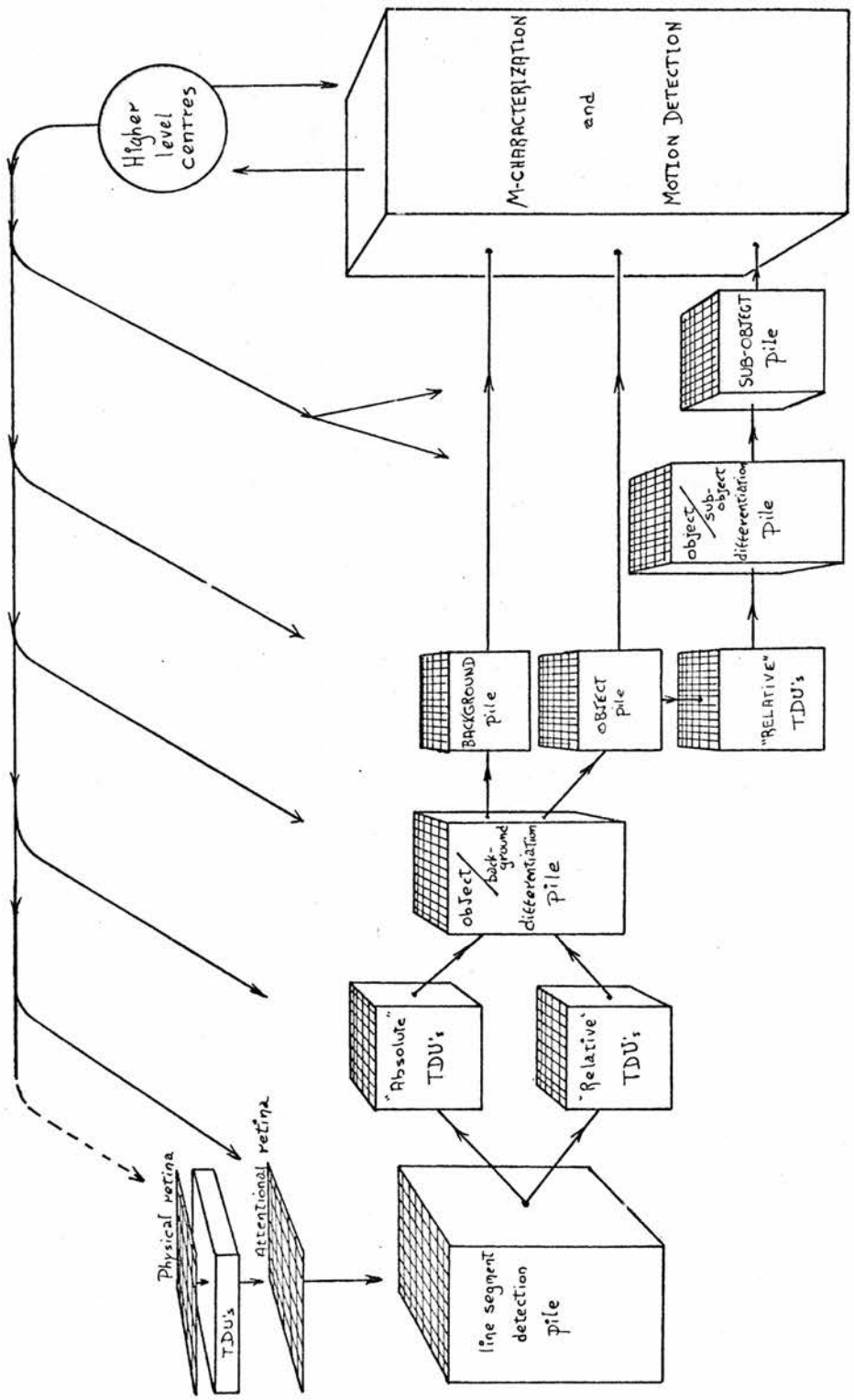


Figure 24. The micro-structure: first part

distinction (the closest line segment to the centre being possibly specified more precisely by the system as being for instance the closest ON-position line segment to the centre); a third one is the choice of the actual criterion on the basis of which the line segments will be sent either to the background or to the object pile; and finally a fourth and a fifth level are equivalent to the second and third level but in the context of the object/sub-object distinction.

V.2 M-characterization and motion detection

V.2.1 Basic intuitive ideas

As we set out to discuss the detailed processes by which the visual object should be M-characterized and by which motion should be computed it might be worth recalling a few very important points concerning the visual object itself. The last section was devoted to preparing and achieving the grouping of a.v.e.'s available from the physical retina into one single visual object. In Part I (Chapter III) it was decided that our system should be made to compute motion on a single object only, descriptive power having to be gained from the diversity of M-characterizing features called in to specify this object. It is only to allow for such a diversity that the system was required to define, besides the visual object itself, two other sets of a.v.e.'s, viz. the background and the sub-object. Since the visual object will in this section be our main concern we want to make sure that the nature of this visual object is completely understood.

First, the visual object does not necessarily represent a single complete physical object in the visual field, i.e. the visual object can represent any part of a physical entity or any group of physical entities lying in the

observed scene (e.g. a flock of birds can easily be the visual object). Secondly, the physical object derived through the first half of our micro-structure cannot bear, at that stage, any other global characteristic than that of the grouping criteria used to generate it. Given a scene where a lorry is moving relative to some background, this means that grouping all line segments making up the lorry on the basis of some particular transistence value yields a group of line segments which are only known by the system as belonging to a "running whole". To find out that this whole is actually a lorry requires higher level skills than the first part of our micro-structure is designed to handle by itself (although it could and should be used in the larger context of those higher level skills). In fact we do not even have to go into such deep semantics as those implied in identifying lorries to find descriptive features which are not directly available from the grouping criteria used to generate the visual object: even the actual motions of this visual object are also left to be derived. Although a moving object like the lorry of our visual example can be identified as a single whole by our system the transistence values which are used as grouping criteria do not say much about the actual motions of the lorry; in the context of field effects with random dots this means that although different groups of moving dots can be adequately recognized as different

running wholes their actual velocities are not made available to the system. In order to reach these velocities, and become aware of the fact that for instance the lorry is undergoing translation and rotation relative to its background or that some set of random dots are moving upwards at such and such a speed, we need M-characterization and its associated motion detection schemes.

The first question is how should the visual object be M-characterized if the system is to detect all useful types of movements in a two-dimensional space.

Looking at the set of line segments making up the visual object at any moment one can easily realise that as a whole this set of line segments can only undergo three different types of rigid motions, any combination of these three types of motion being possible at any moment. The first one is translatory motion, observed when the visual object changes its (global) position; the second one is rotatory motion, observed when the visual object changes its (global) orientation; and the third one is "zoom-lens" motion, or expansion-contraction motion, observed when the visual object changes its (global) size. The first obvious reference for these three types of motion is the retina, as it is in fact this reference which is kept in

the visual object pile's setting of line segments. The first three M-characterizing features to be proposed are therefore (global) position relative to the retina (1), (global) orientation relative to the retina (2), and (global) size relative to the retina (3).

In many cases the system has much more to gain from detecting motion of the visual object relative to its stimulus background than relative to the retina. So the idea is to define a background relative to which the visual object can be M-characterized. Here again position, orientation, and size seem sufficient to cover the range of possible relative rigid motions, so we propose as next three M-characterizing features for the visual object (global) position relative to the background (4), (global) orientation relative to the background (5), and (global) size relative to the background (6).

If the visual object always consisted of a single line segment, the six M-characterizing features would be sufficient to cover the range of its possible motions (the only interesting type of motion lying outside the power of our six M-characterizing features being "curvature motion", observed when the single line segment changes its curvature from moment to moment). But in the overwhelming majority of cases the visual object will consist of a set

of many line segments, thereby allowing systematic changes to occur within the visual object itself and quite outside the reach of global features such as the position, the orientation, and the size of the set as a whole. These changes will be considered as being "shape" changes and will be tentatively accounted for by three new M-characterizing features bearing respectively on the position, the orientation and the size of a well-defined subset of the visual object's set of line segments, expressed relative to the visual object's (global) position, (global) orientation, and (global) size. The idea is therefore to have the system define such a subset of line segments, or sub-object, and to derive this sub-object's position, orientation, and size relative to the object's own position, orientation and size. The new M-characterizing features are therefore sub-object position relative to object position (7), sub-object orientation relative to object orientation (8), and sub-object size relative to object size (9).

Finally, to complete the quick enumeration of M-characterizing features which seem to be required for two-dimensional motion detection at our level of interest, we have to include motion itself, allowing our system to reach at least a second derivative of (featural) distance over time. We therefore propose as last M-characterizing

features for our visual object the speeds of the nine different motions computed on the basis of the nine M-characterizing features already adopted, allowing as many accelerations or decelerations to be computed. This takes the overall number of M-characterizing features to eighteen.

Let us now have a closer look at each one of the proposed M-characterizing features and see what their actual detection implies.

Firstly, in order to derive the visual object's first three M-characterizing features (viz. position, orientation, and size relative to the retina) there is no need to consider any other line segments than those contained in the visual object's pile. Positions, orientations, and sizes of line segments found in this pile are already expressed in purely retinal terms.

To derive the visual object's position relative to the retina at any moment the system only has to use a consistent strategy for raising a particular single local retinal position to the status of global position, the chosen local position having to be somehow representative of the set of positions of line segments lying in the visual object pile at any moment. The strategy which we

propose is one where from each line segment's position on the retina signals are sent along straight paths in all directions, the global position of the set of line segments being the retinal position which is the first one to be crossed by signals coming from all line segments in the set.

Deriving the visual object's orientation relative to the retina is a little more complex. The orientation of a two-dimensional object is not as easily derived as that of a one-dimensional object, i.e. a line segment, and the kinds of strategies required for reducing the "orientational" ambiguity of two-dimensional objects can range from the lowest to the highest level, from raising to a global status the orientation of the longest line segment in the visual object pile to choosing the orientation on the grounds of what "known" physical object the visual object represents. We have already decided that we would not allow the system to wait until the "deep meaning" of the visual object has been identified before triggering motion detection, but we do not want to have such a low level strategy for deriving global orientation that the smallest changes undergone by its local elements fool the system completely. The compromise is to have a "symmetry oriented" strategy where the orientation allowing the most symmetrical description of the set of

line segments making up the object is adopted as global orientation, and where the chosen orientation in cases of ambiguity is the one which lies closest (orientation wise) to the previous moment's chosen orientation. If there is no previous moment's orientation, the closest detected orientation to verticality is chosen. This scheme can easily be made to include "expected" orientations as well as "previous moment" orientations as criteria for choosing the current moment's orientation.

Finally, deriving the visual object's size relative to the retina presents us with a very special problem. In the case of a line segment, it was easy enough to derive size: the length of the line was all that had to be derived. With the visual object, however, we have to be able to deal with two-dimensional size, and we do not want to have to compute areas because of the complexity that would have to be introduced in our computational schemes in order to do so. An alternative way of detecting changes of size involves only considering the end points of all line segments lying in the visual object pile at any moment. The idea here is that any change in the global size of the visual object is bound to affect systematically the length (and thereby the position of the end points) of the local line segments which this visual object consists of. But what can the system do with all these line segments' end

points? The first point to be realised before answering this question is that the rather quantitative nature of size opens up a choice regarding the type of change which we want the system to detect: we can choose to have the system detect either absolute or proportional change in the visual object's value of size relative to the retina. Proportional size indicates how many times bigger or smaller the currently detected size is relative to the size detected at the previous moment, and as far as motion is concerned is much more valuable information than absolute size. The main problem with proportional size is that its computation involves non-linear operations. What we propose is to linearize these operations by using again a pile structure to carry out proportional size detection through a template matching scheme such as the one already used for detecting orientations. This proportional size detection pile will start its analysis from the segments' end points mentioned above, and should yield as output the visual object's proportional size relative to its previous moment's size. The details of how this new pile could operate will be discussed in the next section.

Turning to the problem of deriving the next three M-characterizing features (viz. global position, orientation and size of the visual object relative to some background), the first requirement is to define some frame

of reference lying beyond the retinal frame of reference. This new frame of reference, the background, can be found in the set of line segments lying in what we previously called the background pile. The first thing to do with the background line segments is to characterize them globally with features which will stand as reference for those features of the visual object which have to be derived relative to the background. This of course amounts to giving to the background line segments a global position, a global orientation, and a global size relative to the retina (the retina being of course always the ultimate frame of reference) in the same way as the visual object is provided with such features. So we now have a visual object and a background with their respective global positions, orientations, and sizes relative to the retina. The next and final step consists simply in finding out what shifts are required to bring the background's position, orientation and size (relative to the retina) to a fixed position, a fixed orientation, and a fixed size (e.g. the centre of the retina, verticality, a size ratio of 1), and to actually have the visual object's position, orientation and size relative to the retina undergo those shifts at every moment, thereby yielding the desired position, orientation and size relative to the background. This way of having relative features derived should be highly reminiscent of the way

in which relative positions were derived for the purpose of grouping line segments on the grounds of the transistence values of their relative positions. The basic idea behind the two schemes is obviously very similar, but there is an important difference which has to be introduced in the present scheme: the visual object's global position relative to the background has to be derived in a way which takes into account the orientation and the size of the background. Given two objects lying on the retina, working out the position of one of them relative to the other implies finding how far away and in which direction from the reference object's position the other object's position lies. We see two main ways of doing this. The first way consists in working out the required distance and direction using retinal standards; this would yield what we consider to be a "weak" relative position, and it is the type of relative positioning which was proposed in the last section in our search for adequate running criteria for grouping line segments. The second (much more powerful) way consists in working out the required distance and direction using standards which are inherent to the reference object itself, i.e. using its actual size scale as basis for evaluating the distance and its actual orientation as basis for evaluating the direction; this would yield what we consider to be a "strong" relative position, and it is the type of relative

positioning which we want to use in working out our visual object's position relative to the background.

Moving on to the next three M-characterizing features (viz. global position, orientation, and size of the visual sub-object relative to the visual object) we find ourselves in a very similar situation, where the sub-object is to the object (and vice-versa) what the object was to the background (and vice-versa) in the last situation. The similarity is however not total, the sub-object being a sub-set of the visual object's set of line segments while the visual object and the background consist of disconnected sets of line segments. However partial the similarity might be, it is sufficient for the same computational schemes to be used to derive the three M-characterizing features concerned in both cases. The line segments lying in the sub-object pile therefore only have to be given a global position, a global orientation, and a global size relative to the retina following the usual strategies for doing so, the result of this operation being transformed to yield the sub-object's position, orientation and size relative to the visual object following the same strategy as the one used to transform the visual object's position, orientation, and size relative to the retina into position, orientation, and size relative to the background.

It was decided at the end of Part I (in Chapter III) that the sub-object's position, orientation and size relative to the visual object would be our way of detecting "shape" changes. This decision has some interesting implications. The first one is that shape can only be expressed by relating some local feature of the object (i.e. the position, orientation, or size of the sub-object) to some global one (i.e. the position, orientation or size of the object). This means that our system will never express shape directly in terms of how the local elements of the visual object relate to each other. The second is that since shape is essentially described through local-global (as opposed to local-local) relationships the visual object's global features are of prime importance in defining shape. This means for instance that the representation of the visual object's shape will vary as the chosen global orientation for this visual object varies. In such a system where orientation and shape are totally interdependent any one of the two can be called to determine the other. The third and last is that since in our scheme the system is only allowed to deal with a single sub-object at a time (although this sub-object can be anything from a single line segment to the complete set of line segments making up the visual object) a sequential processing of the visual object's main interesting parts and sub-parts will have to take place for the system to

achieve a complete representation of the visual object's shape.

We believe that our ideas about shape representation can withstand the pressure of a surprising variety of object shapes, including rather complex ones. However, we realised recently that these ideas were not powerful enough to deal with the variety of "shape motions" which the human visual system seems to be able to handle. Although very little work has been done in this direction, a very slight extension of our present shape representation scheme was found to increase considerably the power of the scheme. The idea is that besides having the already explained strategies for deriving the sub-object's position, orientation and size the system would be allowed to use other strategies involving the choice of a position, an orientation, or a size which stand for something precise that all the line segments in the sub-object pile have in common. As the system stands the sub-object's position, orientation, and size are worked out by kinds of "averaging" strategies; what we want to provide is a set of strategies which are not so much concerned with "averaging" as with having specific characteristics of the line segments represented directly through the global position, the global orientation and the global size. Such strategies would be especially well

suited for cases where all the sub-object's local elements bear common characteristics, i.e. cases where the sub-object is "symmetrical" in one way or another. This symmetry will be expressed through the single particular position, and/or orientation, and/or size which represents best the common aspect of all local elements concerned. The power of such a scheme will of course rest on the system's ability to work out symmetries of different kinds. The advantage of using symmetries in a system which can only process one sub-object's position, one sub-object's orientation, and one sub-object's size at a time should be obvious, even though we have not been very explicit about what we mean by "symmetry".

Summing up the main computations proposed so far in this chapter, our system should firstly provide each one of the three sets of line segments found in the background pile, the object pile, and the sub-object pile with a global position, a global orientation, and a global size relative to the retina. The global position, orientation and size given at this point to the visual object's set of line segments should be kept as M-characterizing features. The system should then on the one hand derive and keep as M-characterizing features the visual object's position, orientation, and size relative to the background's position, orientation, and size, and on the other hand

derive and keep as M-characterizing features the sub-object's position, orientation, and size relative to the visual object's position, orientation,and size. This makes a total of nine M-characterizing features derived and leaves us with nine more to derive in order to reach the originally planned total of eighteen M-characterizing features.

Since the nine remaining features are simply the different speeds of the motions computed on the nine frozen M-characterizing features already discussed we have now reached the level of actual motion detection.

Our intuitive ideas about how any motion should be computed have been discussed sufficiently exhaustively (in Chapter IV) to allow us to proceed very quickly through what is required in order to compute the particular motions associated with the nine frozen M-characterizing features discussed so far. Firstly, the nine M-characterizing features' sets of values should conform to a systematic setting along specific dimensions in the system's data structures, and secondly each one of the nine resulting pools of values should be provided with a set of transistence detection units (or TDU's) and an appropriate "travelling OFF network" (or VDU) responsible for detecting the particular feature's associated velocity

at every moment (the nine VDU's of course carrying out their respective velocity detection tasks in parallel). Two interesting ideas can be emphasised in this context. Firstly there should not be any need to find new data structures to express the different possible values of whatever global position or orientation the system happens to be concerned with (this applies to six M-characterizing features out of nine), the data structures proposed (in the last section) for line segments' position and orientation being totally adequate. Secondly, detecting translatory movement is so similar to line segment detection (the former being the running equivalent of the latter) that the line segment detection pile should fit perfectly the processing purposes of at least those VDU's associated with "positional" M-characterizing features (i.e. three out of nine M-characterizing features). Detecting translatory motion in a "line segment detection" type of pile of course offers all the facilities which are provided for detecting the line segments themselves, including the interesting curved line segments detection schemes (yielding "curving" translatory movements in the motion domain).

Once the system has computed the nine motions associated with the nine frozen M-characterizing features, it only remains to treat the respective speeds of these nine

different motions as (running) M-characterizing features to reach our planned total of eighteen M-characterizing features. The motion to be associated with each one of these nine running M-characterizing features will be computed according to the usual principles.

This leaves us with only one aspect of our system to discuss: eye-tracking. There are many reasons for wanting to introduce eye-tracking abilities in any visual system. The two main ones are first to facilitate (or even in many cases "make possible") the frozen analysis of any moving (physical) object by immobilizing it relative to the physical retina, and second to facilitate certain aspects of the running analysis of (physical) objects moving relative to the (physical) object being eye-tracked by making these objects' motions relative to the tracked object directly available from their motions relative to the physical retina.

To be consistent with the generality of the reasons given above we should provide our system with three different types of eye-tracking abilities, one for each type of detected motion relative to the retina, i.e. translatory motion, rotatory motion, and "size change" motion. Ideally, each of them should be accounted for by some appropriate tracking system, but the problem involved in

setting up tracking schemes for changes of size (which requires a retina with adjustable size, or some zoom-lens system) and changes of orientation (which requires a retina that can rotate clockwise and anticlockwise), without mentioning the problem involved by having many tracking systems in operation at the same time, forced us to consider the tracking of translatory motions only.

The problem with eye-tracking is that by involving a motion of the retina relative to the organism a situation is created where our motion analysis system becomes totally incapable of providing velocities relative to the organism. Our motion detection system was designed assuming that motions relative to the retina are equivalent to motions relative to the organism ("organism" referring to the physical system to which the eye is linked), and introducing an eye-tracking system breaks the equivalence between the two frames of reference. Considering that for both purposes of driving the eye-tracking itself and of adapting to some environment on the basis of the evidence provided by the visual system motion relative to itself is what counts for the organism, what we need is some means of transforming velocities relative to the retina which are affected by eye-tracking into velocities relative to the organism. So although it is not itself involved in the strict detection of

velocities, eye-tracking calls for quite a substantial re-interpretation of some of the primary velocity information provided by the actual motion detection schemes. For this reason we decided to consider the eventual eye-tracking system (with its associated schemes for re-interpreting actual velocities detected relative to the retina) as a logically distinct motion interpreting system, and we decided to stress its dependence on actual motion detection by calling it the "Secondary system" of visual motion detection, the whole of the motion detection schemes involving the eighteen M-characterizing features discussed in the above paragraphs becoming the "Primary system" of visual motion detection.

V.2.2 Nerve net embodiments

The first step towards the complete M-characterization of the visual object is to provide each one of the three sets of line segments found respectively in the background pile, the object pile, and the sub-object pile with a global position, a global orientation and a global size. For the sake of clarity let us assume that the computation of the three features for each of the three sets of line segments will take place in separate piles. The first move is to transfer all the line segments from each of the three starting piles (viz. the background pile, the object pile, and the sub-object pile) into each of the three piles respectively designed for computing global position, global orientation, and global size for any set of line segments. Altogether there will be nine new piles to be loaded from the three starting piles: three global position piles, three global orientation piles, and three global size piles (cf. Figure 29 to see how these nine piles fit in the overall system).

Since the three global position piles and the three global orientation piles will be standard twisted piles, transferring line segments to them is a rather straightforward task: the single cells standing for the particular line segments stored in the original piles

simply get their values transferred to the corresponding cells in the new piles. Once all line segments have been loaded in each pile the computation of global positions and orientations can start.

The strategy proposed in the last section to compute global position can be implemented within each of the three global position piles in the following way. First, each occupied cell is mapped vertically through the whole pile leaving a mark in each of the cells encountered as successive layers are crossed (one cell per layer). From each marked cell (including the originally occupied cells) a signal is sent (in the plane of the layer) in the two possible directions through the layer in which the cell is sitting, this signal leaving a mark in every cell which it crosses on its way. This is the scheme by which signals from line segments' "positions" are sent in all directions, each layer of the pile allowing for two "travelling" directions (180 degrees apart). The way in which the system will find the first single retinal position where "travelling" signals from all line segments in the pile pass is by constantly checking on the number of overlapping cells (vertically linked to every retinal position field) which have been crossed by some signal travelling in the plane of the layers; as soon as this equals the number of line segments in the piles for any

given set of overlapping cells the retinal position which sits on top of this set is chosen as global position.

Now the implementation of a global orientation detection strategy of the type proposed in the last section can be carried out within each one of the three global orientation piles in the following way. First, through the single cell which represents each one in the pile, all line segments are mapped (vertically) from their respective layers of origin onto every other layer of the pile. The content of each cell onto which they are mapped is no longer devoted to size but is turned into a means of storing the vertical distance (expressed in terms of a number of layers) by which the line segment specified by the given cell stands away from its original layer. What we have at the end of such a mapping process is a pile where all line segments which were originally sitting in their respective layers are now represented on each one of the pile's layers, the contents of the cells in each of these layers telling how far they stand from their original layer. By taking any layer of the pile one can then easily tell if the orientation for which this layer stands allows any symmetry to stand out. Orientationally symmetrical line segments can obviously be detected in such a context by the fact that on the layer which allows this symmetry the line segments concerned will be at the

same "vertical distance" from their original layers. The scheme can be improved in many ways, by introducing for instance "horizontal" distance between the line segments on every layer as a basis for working out positional symmetry, or by introducing weights which favour certain orientations (e.g. verticality) or certain orientational relations (e.g. orthogonality), these weights being possibly optional (e.g. to suit expectation). Whatever the final set of criteria turns out to be, every layer of the pile is assessed according to the chosen criteria and the "winning" layer's orientation becomes the set of line segments' global orientation.

As far as the three global size detection piles are concerned, the first problem to be tackled is that of designing the piles themselves, the usual twisted pile being totally inadequate for either detecting or storing the proportional two-dimensional sizes which were proposed in the last section as object of global size detection. The proportional size detection pile which we propose consists of a stack of two-dimensional arrays (i.e. a simple non-twisted pile) where the middle layer is the usual (16X16) array of retinal position cells but where a set of layers of decreasing scaling extends the pile upwards and a set of layers of increasing scaling extends the pile downwards. The changing of scale from one layer

to the next should be such that any two layers with a given proportional difference in scale (e.g. one having units twice as large as the other) are separated by the same number of other layers as can be found between any other two layers which have the same proportional difference in scale (i.e. something similar to logarithmic variation). This means that linear distances through this pile, in terms of number of layers, stands for proportions of sizes.

The way in which actual size proportions can be computed in our new pile is as follows. Firstly the line segments for which a global proportional size has to be found are loaded in the middle layer of the pile (the one which is set according to the retinal scaling) in terms of the two end-points of each line segment. This means that each end-point of each line segment is associated with a particular cell of the layer. In order to make sure that changes in proportional size are computed independently of changes in position or orientation, the group of line segments transferred to the proportional size pile should have previously been given both a global position and a global orientation and should have been shifted and rotated in order to be mapped upright with its global position in the centre of the proportional size pile's middle layer. Once the pile has been loaded every

occupied cell in the middle layer is given a pair of coordinates relative to the centre of the layer (0,0). Then all cells corresponding to these coordinates in all other layers of the pile are marked. Since the scaling is different for every layer those cells which have the same coordinates in different layers do not overlap each other (vertically) in the pile. When the next moment comes the new set of occupied cells in the midle layer triggers signals which are sent straight up and straight down through the pile looking for marked cells. The layer in which the greatest number of marked cells are found during these vertical journeys is the winning layer and its (vertical) distance from the previous moment's winning layer specifies the proportional change in size of the group of line segments. The currently occupied cells of the middle layer are used to mark the other layers' cells for the next moment's analysis and the whole process starts all over again.

It is very easy to introduce into such a scheme a specificity to changes of proportional size along either the vertical or the horizontal axis, but we will not go into this for the time being.

Out of the nine features which the system is now equipped to detect three should be used directly as

M-characterizing features, i.e. global position, orientation and size (relative to the retina) of the visual object's set of line segments. The next step, which should lead us to the remaining six frozen M-characterizing features, consists in using on the one hand the background's detected values of position, orientation, and size relative to the retina as basis for transforming the visual object's detected values of these features relative to the retina into values of the same features relative to the background, and in using on the other hand the visual object's detected values of the three features relative to the retina as basis for transforming the sub-object's detected values of the three features relative to the retina into values of the same three features relative to the visual object. The transformation process will be the same for the two levels of relative representation (i.e. the object/background level and the sub-object/object level) and it can be carried out as follows.

For a start we should provide the system with two new sets of piles, one for each level of relative representation, each set consisting of a relative global position pile, a relative global orientation pile, and a relative (proportional) global size pile (cf. Figure 29 to see how these six new piles fit in the overall system). It is in

these piles that the transformation of values of features relative to the retina into values of features relative to the particular sets of line segments will be carried out.

Global position can be dealt with in the following way. Firstly the global position value relative to the retina which is to be transformed is transferred into the relative global position pile in exactly the same location as the one it was occupying in its original pile. This of course implies that the relative global position pile should be a standard twisted pile. In fact the relative global position pile will have to be a combination of a twisted pile and of a proportional size pile in order to derive what was called in the last section a "strong" relative position, the requirement in this case being that both the global orientation and the global size of the reference set of line segments should be taken into account in deriving relative global positions. Once the global position to be transformed has been transferred into the twisted pile part of the relative global position pile it is (1) mapped vertically onto the layer which corresponds to the global orientation of the reference set of line segments, and (2) shifted on this layer in a way which would bring the global position of the reference set of line segments to the centre of the layer (0,0). This yields a relative global position where both the position

and the orientation of the reference set of line segments have been taken into account. In order to obtain a relative global position which also takes into account the (proportional) size of the reference set of line segments the system only has to transfer the relative global position just obtained into the middle layer of the (proportional) size pile part of the relative global position pile and to map it vertically onto the layer corresponding to the current proportional size of the reference set of line segments. This will yield the desired (strong) relative global position.

Relative global orientations can be obtained in the following way. Firstly the "global orientation relative to the retina" which is to be transformed is transferred into the relative global orientation pile. This pile needs only to be a single column of cells, every cell standing for a layer of the original twisted pile. Then the system only has to shift the "global orientation relative to the retina" (represented by a 1 in one of the column's cells) in a way which would bring the global orientation of the reference set of line segments to the column's cell which stands for verticality (or 0 degrees). This yields the desired relative global orientation.

Finally relative global (proportional) sizes can be

obtained in a very similar way. Firstly the "global (proportional) size relative to the retina" is transferred into the relative global (proportional) size pile. This pile needs only to be a single column of cells, each cell standing for a layer of the original proportional size pile. Then the system only has to shift the "global (proportional) size relative to the retina" (represented by a 1 in one of the column's cells) in a way which would bring the global (proportional) size of the reference set of line segments to the middle cell of the column (i.e. the one standing for proportion 1/1). This yields the desired relative global (proportional) size.

This makes altogether nine piles containing current values of M-characterizing features: three specifying the global position, orientation, and size of the visual object relative to the retina, three specifying the global position, orientation, and size of the visual object relative to the background, and three specifying the global position, orientation, and size of the sub-object relative to the visual object. All is now ready for motion detection.

Once again, for the sake of clarity, we will use a new set of piles for the description of how the system should deal with the next computational task, i.e. motion detection.

This means that for each one of the nine M-characterizing feature piles a specific motion detection pile will be provided (cf. Figure 29 to see how these nine piles fit in the overall system).

The motion detection piles associated with the three global position piles will be standard twisted piles. Motions detected within these piles will of course be translatory motions and they will be detected in the following way. Firstly the single global position which is detected at every moment within each one of the three global position piles will be loaded into each one of the three associated translation detection piles by being plotted on the top layer of the pile (i.e. the vertical or 0 degree layer). This layer will indeed stand as pool of values of the type of position which the particular pile is meant to deal with. In order to conform to the general principles of motion detection discussed in Chapter IV this pool of values has to be provided with an associated pool of TDU's so that the transistence value of every position value is made available to the system at every moment. Now the travelling OFF network (or VDU) which has to be provided to allow detected OFF transistence values to search for detected ON transistence values in all possible directions will be provided by the twisted pile itself on top of which the pool of position

values is sitting. The way in which the pile will be used as VDU is as follows. Whenever an OFF signal is triggered by one of the TDU's sitting on top of the pile this signal is mapped vertically through the whole pile, and for every layer encountered a signal is sent in the plane of the layer in the two directions (+ and −) allowed by the particular layer's orientation, starting from the layer's cell which has been touched by the OFF signal. This means, if one considers the whole pile, that signals are sent in twice as many different directions as there are layers in the pile. On the other hand whenever an ON signal is triggered by one of the TDU's this signal is simply mapped vertically through the whole pile, leaving a mark in every cell which is crossed as the signal goes from one layer to the next. Now whenever both an OFF signal and an ON signal are triggered within the same moment, i.e. when some translatory motion is to be detected, there is in principle only one layer which can allow for an OFF signal to meet a cell marked by the ON signal, and this is the layer whose orientation corresponds to the direction of the translatory motion to be detected. The distance between the position going OFF and the position going ON is easily available from the distance, in terms of cells crossed, travelled by the OFF signal on the particular layer where the ON signal is found. We therefore propose a scheme where OFF signals

travelling on the different layers of the pile keep track of the distance which they travel as they search for cells which are marked by an ON signal; when such a signal is found the layer on which this happens together with the "sign" (+ or −) of the signal having made the finding are taken to specify the direction of translatory motion, and the distance travelled by this same signal is taken to specify the speed of translatory motion. The similarity between this translatory motion detection scheme and the frozen line segment detection scheme discussed previously should be obvious.

Now the three motion detection piles associated with the three global orientation (M-characterization) piles will simply be single columns of cells of the type already used for two out of the three M-characterization piles. Rotations can be computed in such piles or columns through a rather trivial setup consisting on the one hand of a set of TDU's associated with the pool of values represented by the set of cells in each column (one TDU per cell), and on the other hand of a straightforward travelling OFF network (or VDU) involving two "travelling OFF lines" (i.e. two detectable directions of motion), one allowing OFF signals to travel upwards in the column (i.e. an "anti-clockwise" line) and the other allowing OFF signals to travel downwards through the same column (i.e. a "clock-wise

line). Each column will be loaded at every moment with the single global orientation detected within the M-characterization pile with which it is associated. The loading process will simply consist in putting a "1" into the column's cell corresponding to the detected global orientation and a "0" in every other cell of the column. A quick look back at Figures 17 and 18, in Chapter IV, might help one to visualise the simple rotation detection scheme which is proposed here.

Finally the three motion detection piles associated with the three global proportional size (M-characterization) piles will also be single columns of cells similar to those already used for two out of the three M-characterization piles. "Size motion" can be computed in the same straightforward way as rotation since only two different directions of motion (viz. expansion and contraction) have to be considered. Each one of the three "size motion" columns will be provided with a set of TDU's (one TDU per cell) which will be linked to a VDU consisting of two "travelling OFF lines", one allowing OFF signals to travel upwards in the column (i.e. a "contraction" line), and the other one allowing OFF signals to travel downwards through the same column (i.e. an "expansion" line). Each column will be loaded at every moment with the single global (proportional) size detected

by the M-characterization pile with which it is associated, and this loading will be similar to the one described in the above paragraph for the case of global orientation.

Once the nine motion detection piles described above have computed, in parallel, their respective velocities the only motions left to be computed by the system are those of the actual speeds of the velocities just detected. It was decided in the last section that accelerations and decelerations within the nine different basic velocities should be computed by our system. However, before the speeds of all nine different velocities obtained so far are considered as M-characterizing features whose motions have to be computed there is one computed velocity which must be modified. The velocity concerned is the translatory velocity of the visual object relative to the retina and the modification consists in transforming this detected velocity relative to the retina into a velocity relative to the whole organism within which the eye is sitting. We are of course entering here the context of the Secondary System of visual motion detection referred to in the last section in relation to our desire to provide the system with eye-tracking abilities.

The general idea behind the secondary system is that in

order to perform efficient eyetracking, and in order to allow an optimum awareness within the visual system while the eyetracking is under way, we need a system which computes velocities relative to the organism whenever some eyetracking is going on.

Our first version of the secondary system worked in the following way. When some object is to be tracked by the eye, this object's velocity relative to the retina (as worked out by the primary system) is sent to the secondary system. In the "initialising" phase of the eyetracking the task of the secondary system is quite straightforward: since no tracking was taking place just before, the object's velocity relative to the retina is equivalent to its velocity relative to the organism, and the secondary system only has to put the label "velocity relative to the organism" on its input to transform it into the desired output. Now this output is sent (1) to other parts of the visual system for further analysis; (2) to the oculomotor system where the eyetracking is triggered off; and (3) back to the secondary system itself as an input for the next moment. When the next moment comes, since the eyetracking is on the way, the object's velocity relative to the retina is no longer equivalent to its velocity relative to the organism, so that this time the secondary system has got a little more work to do to get its output

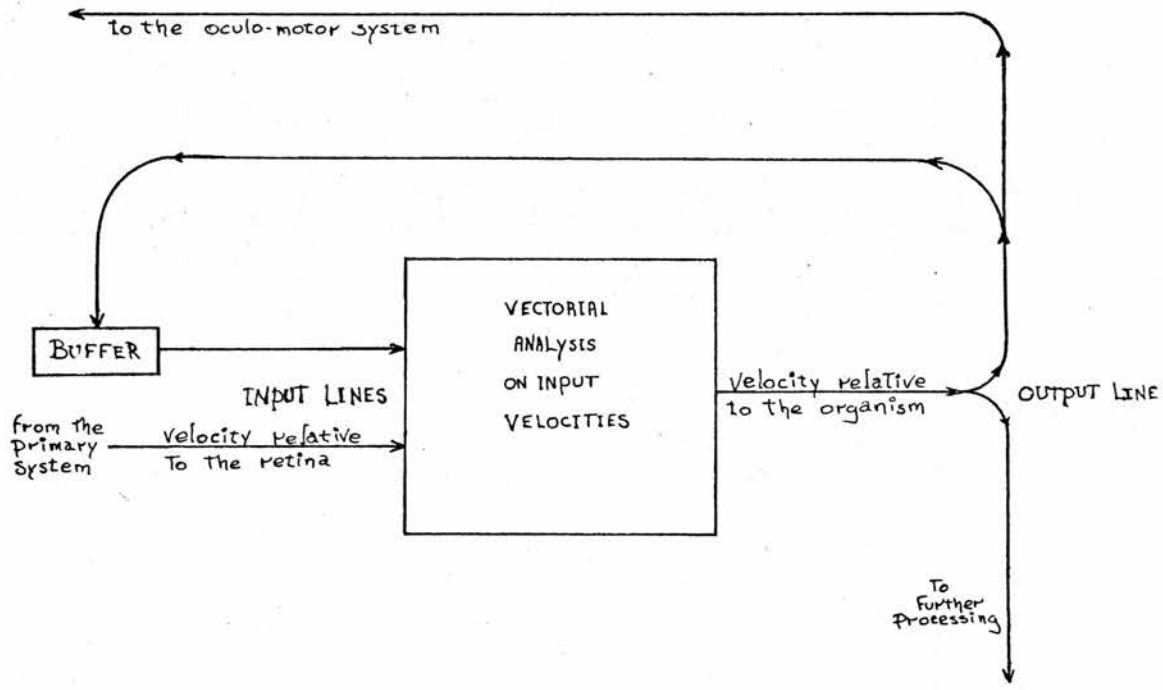


FIGURE 25. The secondary system: first version.

computed. The way in which the object's velocity relative to the organism is computed in this case is that the secondary system carries out a vectorial analysis on its two input velocities, namely the object's velocity relative to the retina as provided by the primary system, and the eyetracking velocity as provided by the secondary system itself through its own latest output (third item in the list above). The output generated in this way by the secondary system is then sent to the same three destinations as in the "initialising" phase, and the

procedure that we have just finished describing is repeated until either the object can no longer be tracked or the visual system dismisses it as the focus of attention. It might be worth noting that the secondary system can be considered at all times as the vectorial analysis system that we just described, the "initialising" phase being a case where one of the two input velocities (namely the eyetracking velocity) is a null velocity. The schema in Figure 25 might advantageously express what we tried to describe verbally in the above paragraph.

This secondary system allowed sustained eyetracking in cases of nonuniform motion as well as in cases of uniform motion. However, we noticed that in the case of tracking an object in uniform motion, since the tracked object is at all times kept "still" relative to the retina, the primary system is only involved in computing immobility and the secondary system is reduced to carrying out vectorial analysis involving only one non-null velocity, namely the eyetracking velocity. This means that trivial computations are monopolising both the primary system and the secondary system (since they can only work on a single object's velocity at a time, however trivial the computations are) whenever the tracked object is in uniform motion. This is fair enough if one realises that in most cases one cannot tell when the motion will cease

to be uniform and that, when it does, the full power of both systems will be needed to make the necessary corrections. Nevertheless, we came to the conclusion that we ought to provide the secondary system with an auxiliary system which could take care of the "simple" computing required when the tracked object is in uniform motion; this would leave the visual system free to investigate other moving objects in the field of view while automatically tracking the object it was initially interested in. We therefore decided to split the secondary system into two distinct parts: the main secondary system (MSS), which is a structural replica of the former secondary system, and the auxiliary secondary system (ASS), which is the new structure for handling the eyetracking of objects in uniform motion. This new development was a complication, but the power of the secondary system was considerably increased.

Here is how the new secondary system works. The information provided to the secondary system by the primary system is divided into two groups: the null velocities (ON and STILL transistence values computed by some TDU) which are sent to the ASS, and the non-null velocities (computed by some VDU) which are sent to the MSS. The MSS is used to "initialise" the ASS in the following way. Whenever the visual system decides to

track an object, this object's velocity (as computed by the primary system) is sent to the MSS (it has to be a non-null velocity or else no tracking would be required). The MSS then carries out its vectorial analysis and sends the result (1) to other parts of the visual system for further analysis; (2) back to itself as an input for the next moment; and (3) to the ASS. The ASS swallows the input and, without bringing any alteration to it, (1)sends it to the oculomotor system where the eyetracking is triggered off; and (2) feeds it back to itself as input for the next moment. Apart from its own output the moment before (which we will call the "local" input), the ASS can receive only one of the two possible inputs (which we will call the "foreign" inputs) at every moment: it is either the output of the MSS (when a new tracking velocity is required), or a null-velocity signal from the primary system (when the motion of the tracked object remains uniform). When the foreign input is the MSS's output, then the ASS ignores the local input (from its own latest output) and goes through the same routine as it did when "initialised" ((1) and (2) above). However, when the foreign input is the null-velocity signal from the primary system, then the ASS takes the local input and (1) sends it up as a command to the oculomotor system; (2) sends it to the other parts of the visual system for further processing; and (3) feeds it back to itself as a local

input for the next moment. As long as the tracked object remains in uniform motion, the ASS can handle it very well on its own by going through the loop we just described. A representation of the main features of the new secondary system is given in Figure 26.

An interesting point is that this new secondary system gives our system the ability to "shift" the eyetracking from one object moving at a given velocity to a different object moving at a different velocity without having to break the eyetracking in the process. This is actually done in two steps: first the computational power of both the primary system's VDU and the MSS are transferred from the currently tracked object to some other moving object in the field of view, leaving to the ASS the task of handling the eyetracking itself. This transfer is achieved by what we called earlier an attentional saccade. Then the second step consists of transferring the eyetracking itself by "re-initialising" the ASS with the velocity of the new object (as worked out by the MSS after completion of the first step). This transfer will generally be accompanied by a physical saccade for reasons that will be made clearer in the paragraphs that follow. It is important to realise here that if the tracked object's velocity happens to change <u>after</u> the first step has been completed, and <u>before</u> the second one is

completed, then the eye-tracking breaks down (since the ASS alone can only cope with uniform motion). But it is also important to realise that this critical inter-step period can be made very short indeed.

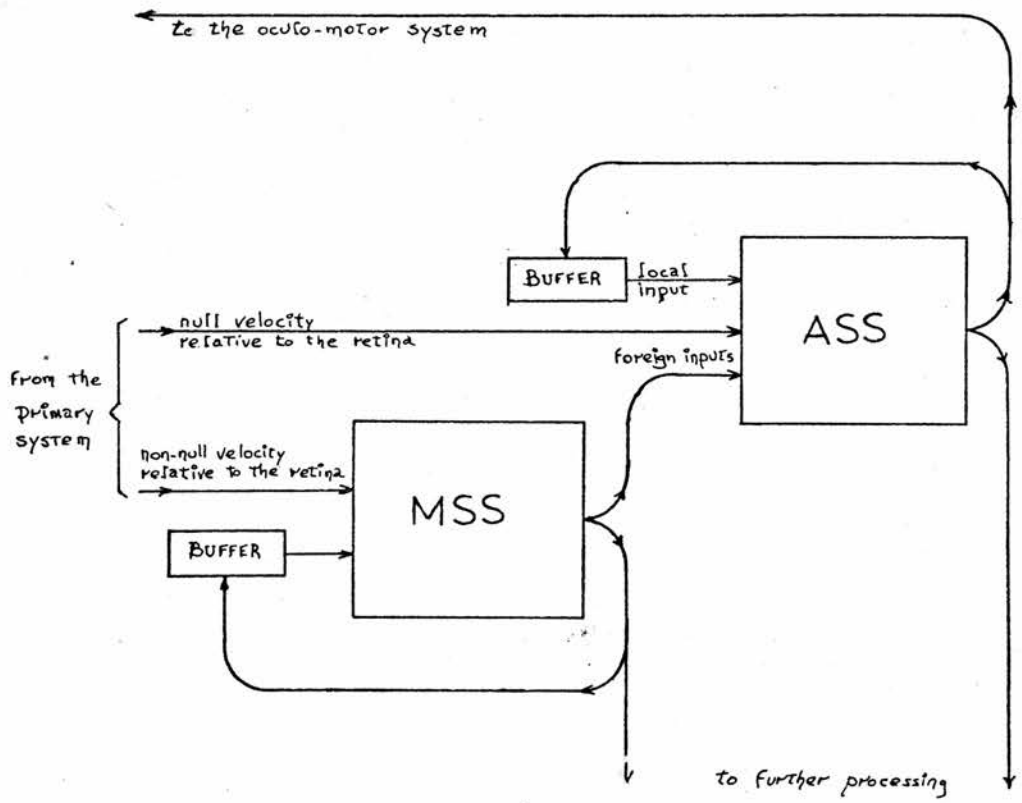


FIGURE 26. The secondary system: second version.

Now this business of changing the eyetracking velocity, whether it appears in the context of tracking a single object in nonuniform motion or in the context of transferring the eyetracking from one object to another, raises an important problem which we have not yet dealt

with. When we take a closer look at how our present system deals with changes of eyetracking velocity, we realise that these changes are inevitably brought one "moment" after the object to be tracked has undergone the change in velocity (for the obvious reason that the change has to be detected before the system can start coping with it). The consequence of this delay is that the eye either loses or gains ground on the object (depending on the type of change involved) every time a change of velocity occurs, and, since the visual field of the eye is a limited one, a succession of such changes might progressively "push" the object out of the visual field. We concluded from this that, whenever a change of eyetracking velocity is required to match the tracked object's velocity, an "extra" motion is also required to catch up with the object. This "extra" motion can in fact be carried out simply as a physical saccade when the new eyetracking velocity starts operating. So along these lines we decided to allow saccades within the eyetracking process itself: the commands for saccades would simply be combined with the commands for tracking whenever required. This new setup allowed the visual system to keep the tracked object more or less in the same spot on the retina throughout the whole tracking. To optimise the situation we also decided to make it a rule that the tracked object should be kept in the most central region of the retina.

There is only one problem left to be tackled before closing the discussion of our secondary system, designing precise nerve nets which can carry the load of the processing for which the MSS is responsible.

The first point is that we want the MSS' input velocities to be expressed in terms of pairs of orthogonal velocity components, for two main reasons. The first is that the vectorial analysis to be carried out by the MSS requires such a format, and the second one is that the eye-tracking velocity command, which finds its origin in the MSS's output velocity, also requires the two-components format. The reason why the oculo-motor system requires velocities split in two components is that we think of it as a system which can drive the eye in many different directions through different combinations of activations of only two sets of effectors, one possibly representing the vertical component, and the other one representing the horizontal component. Such a system is obviously much more economic than having a particular set of effectors for each different direction in which we want the eye to be able to move.

Now we only have to worry about one of the MSS' two input velocities in our effort to express velocities in terms of orthogonal components, one of them being already expressed

in such terms due to the fact that it consists of the MSS' own output at the previous moment. The input velocity about which we have to do something is the one which comes from the primary system, i.e. the visual object's (translatory) velocity relative to the retina . One way to express this velocity in terms of two orthogonal components would be to take it as it comes out of the VDU part of its specific motion detection pile and split it into the desired components using another standard twisted pile. But the question is why bother computing the single velocity in the first place if we are going to split it anyway; is it not possible to compute the required velocity components directly and postpone the computation of the single velocity until the vectorial analysis has provided us with means of expressing it relative to the organism? This is quite possible and we propose to do it in the following way.

Firstly everything happening in the motion detection pile concerned with the visual object's translatory velocity relative to the retina will remain as described above up to the detection of OFF and ON position values on the top layer of the pile. Then instead of being sent down through the rest of the pile for actual velocity detection the OFF and ON signals will be kept on the top layer of the pile, and it is on this vertical (or 0 degree) layer

that the system will be made to derive the required orthogonal components of the visual object's translatory velocity relative to the retina. (Any other single layer could of course have been chosen as reference for providing the orthogonal components.) The actual components will be found by having both the OFF and the ON position cells send signals in the four possible directions allowed by the chosen layer: along the row to the right (h+), along the row to th left (h−), upwards along the column (v+), and downwards along the column (v−). The signals coming from the OFF positions will keep track of the distance travelled as they go, all signals leaving a mark in the cells crossed on their way. <u>There will always be two and only two cells which will each be crossed by two signals</u>. One of these two cells will be crossed by one "OFF position column" signal and one "ON position row" signal: the distance travelled by the "OFF position column" signal before reaching this cell together with the sign (+ or −) of this signal will specify the vertical component of the visual object's translatory velocity relative to the retina. The other of the two cells will be crossed by one "ON position column" signal and one "OFF position row" signal: the distance travelled by the "OFF position row" signal before reaching this cell together with the sign (+ or −) of this signal will specify the horizontal component of the visual object's

translatory velocity relative to the retina. Figure 27 shows an example of how this "splitting" procedure is carried out.

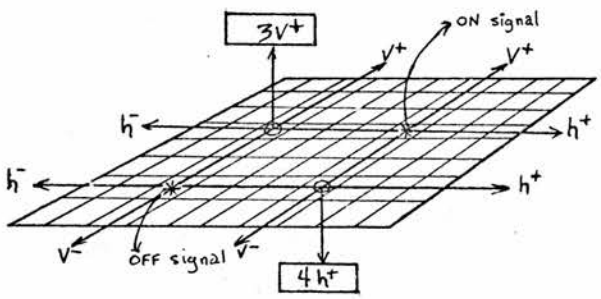


FIGURE 27. The "splitting" step.

Once the velocity components have been worked out in the above described way they are sent out of the pile to the MSS itself where the vectorial analysis takes place. This vectorial analysis, involving on the one hand the two velocity components of the visual object's translatory velocity relative to the retina and on the other hand the two velocity components of the eye-tracking velocity, should yield the two velocity components of the visual object's translatory velocity relative to the organism. It is achieved simply by having corresponding components inhibiting each other in the case of opposite directions,

and being summed together in the case of similar directions. Figure 28 shows how this can be done in terms of explicit nerve nets.

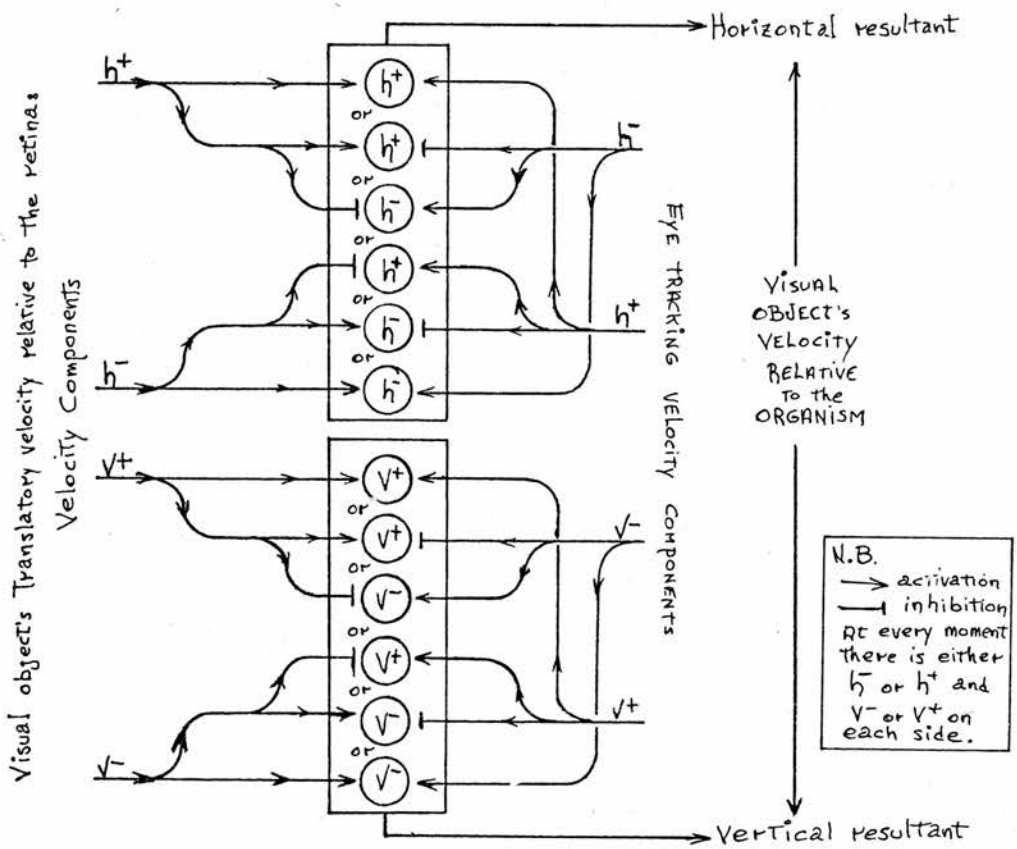


FIGURE 28. The vectorial analysis.

The result of the MSS' vectorial analysis, i.e. the velocity components of the visual object's translatory velocity relative to the organism, can then be sent to the ASS (if the system wishes it so) and fed back to the MSS itself for the next moment's vectorial analysis, but this result is also sent back to the top layer of the motion

detection pile where the original velocity components relative to the retina were obtained. This time however the reverse process is undergone: the velocity components obtained through the MSS' vectorial analysis are used to plot the new position of the ON signal on the top layer of the pile, and when this is done both the OFF and the ON positions are mapped down the pile and a single velocity is derived in the standard way. The motion detection pile in which all this is done therefore becomes specific to the visual object's translatory velocity relative to the organism.

Now that the visual object's translatory motion relative to the retina has been given the wider frame of reference of the whole organism we can come back to our concern for treating the respective speeds of the nine velocities obtained so far as M-characterizing features which should be analysed motionwise. Computing directions and rates of changes of speed amounts to detecting accelerations and decelerations. This can easily be done for each of our nine different types of velocities by providing each with a pile consisting of a column of cells where each cell stands for one given speed and where the speeds represented in the successive cells are organised in quantitative order (decreasing or increasing). Each such pile (or column) is then provided with an associated set

of TDU's and a VDU consisting of two travelling OFF lines, one going in the direction of increasing speed values in the column (i.e. an "acceleration" detection line), and one going in the other direction (i.e. a "deceleration" detection line). Notice that this motion detection setup is designed to detect changes of "absolute" speeds, not changes of proportional speeds. It could well be a better idea to treat changes of speed in the same way as changes in size (the only other "quantitative" M-characterizing feature in our system) were treated.

Figure 29 summarizes the different stages of the processing taking place in this second half of the system's micro-structure. Notice that here again "channels" have been introduced to travel back and forth between precise process and data structures on the one hand and some so-called higher level centres on the other. These channels mean that the system can be provided with information coming from higher level centres as well as providing these centres with some information. We realise here that the interaction between the system and higher level centres is, to say the least, expressed in a rather sketchy way in the schema of Figure 29. However, we want to stress that our discussions about how global orientations can be derived and about how the Secondary System can be made to share its different abilities,

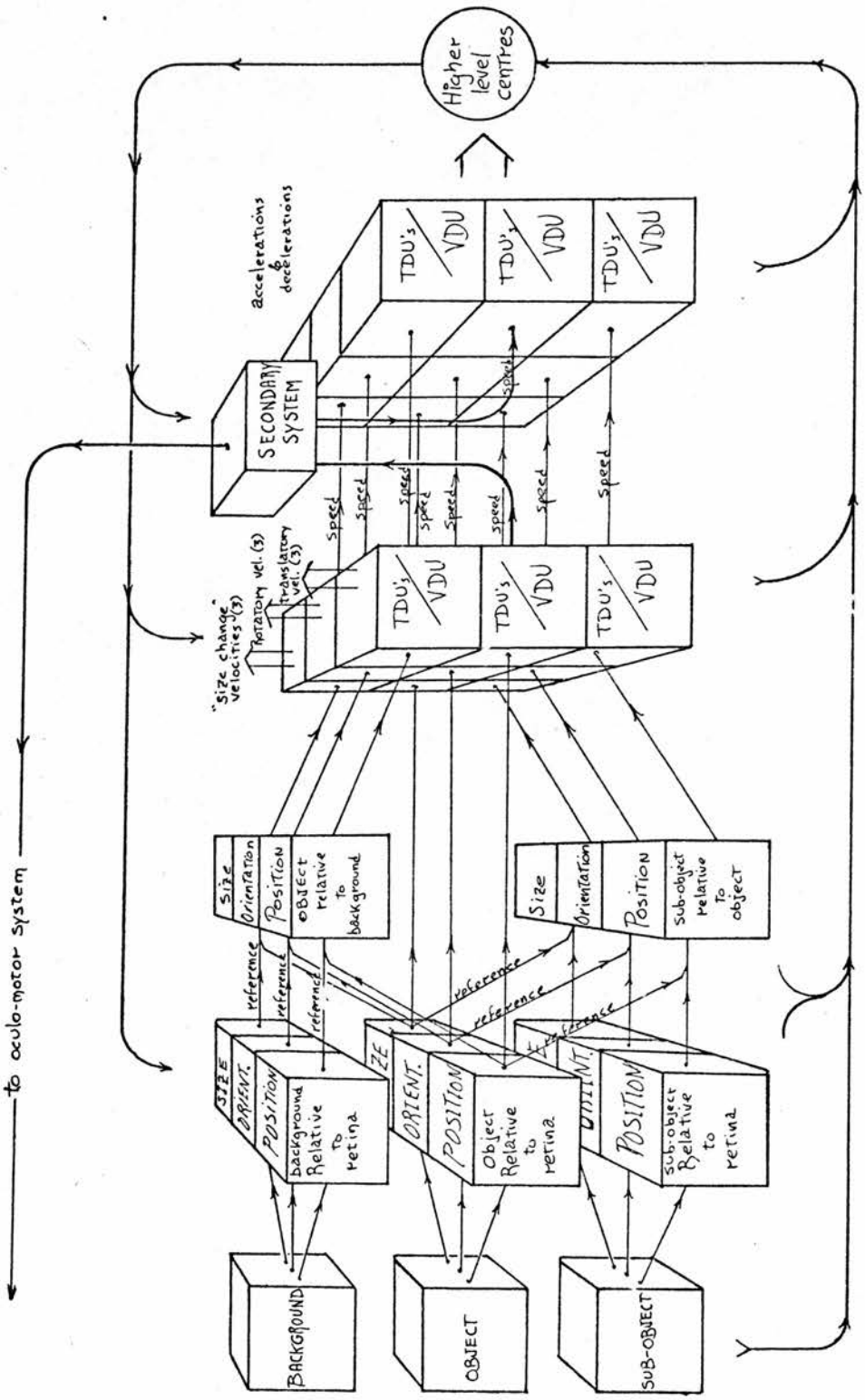


Figure 29. The micro-structure: second part

together with the obvious fact that some selection has to be made regarding which of the absolute or the relative motions of the visual object should be considered in the current moment, are quite sufficient to justify the few extra arrows added to the schema. However, the fact that the emphasis has been placed on bottom-up, or "stimulus driven", processes in the design of the system so far deserves some comments.

For a start it should be made perfectly clear firstly that we fully acknowledge the need for model-driven analyses in a visual system performing at a human level of sophistication and secondly that we do not contest the fact that we have up to now been mainly concerned with designing stimulus-driven analyses. We happen to believe that the two types of processes are complementary aspects of complex visual systems and we intend to eventually provide our stimulus-driven motion detection system with the model-driven counter-part required to reach a human level of visual performance. Before discussing why our emphasis was first placed on the stimulus-driven aspect of visual processes we will try to be a little more explicit about what essentially differentiates the two types of processes.

The distinction between model-driven analyses and

stimulus-driven ones can by no means be said to be well-defined. However most people seem to agree on one point which can be considered to be the main reason behind the distinction, and this point is made rather clearly by Gregory (1972) when he says:

> Current sensory data cannot be sufficient for perception or control of behaviour: it must select relevant facts and generalisations from the past, rather than control behaviour directly from present stimuli.

Stimulus-driven analyses are of course those which bear on "current sensory data", while model-driven ones are those which are based on selections of "relevant facts and generalisations from the past". It is quite obvious that our motion detection system, apart from the provisions made for allowing higher level interventions, mainly involves visual analyses of the first category although it is very interesting to notice that <u>stictly speaking</u> motion detection reaches beyond "current sensory data" and can even be argued (in a very restricted sense) to be involved in "selecting relevant facts and generalisations from the past". Of course the "past" which Gregory refers to is much more remote than the previous moment of analysis and the "generalisations" which he refers to are of a much higher level nature than grouping and M-characterization processes carried out by our system.

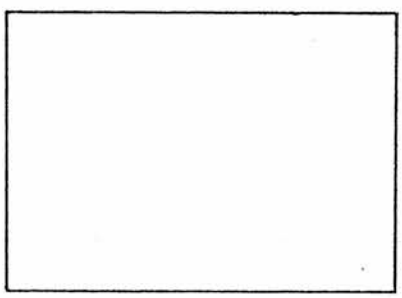
Now the realisation that stimulus-driven processes are not sufficient to account for the richness and flexibility of human vision should not overshadow the fact that they are nevertheless necessary to account for them. After all the organism has to adapt to the environment within which it is currently behaving, not the environment within which it expects to be behaving. This brings us back to the fact that both stimulus-driven and model-driven processes are required and to the question of deciding which one of the two should be tackled first in designing a visual system.

The ideal attitude is of course to tackle both at the same time. Unfortunately this turned out to be an impossible task (for us). The best we could do was to choose one aspect of the processing, i.e. stimulus-driven analyses, and to try to take into account as many high level requirements as possible, including openings in our system for eventual model-driven interventions or manipulations. The reason why we chose to start by placing the emphasis on the stimulus-driven aspect of our motion detection system is simply that stimulus-driven analyses are the basis of model-driven ones, not vice-versa. We believe stimulus-driven analyses to be logical and developmental prerequisites for model-driven processes, and we therefore believe that the road leading to the latter has to go through the former. Indeed how could anyone design

model-driven processes before knowing what stimulus-driven processes are capable of? This might all sound obvious, but it seems that an overwhelming tide of concern for top-down processes has swept over the entire field of A.I. in the recent years, making concerns for stimulus-driven processes look like naive or useless battles against mere technical details. We do agree that the higher level problems which model-driven processes allow to investigate are more exciting, but let us repeat that we believe that the road to these higher level problems should pass through the dryer land of "technical details" if it is to lead to success, and we believe that we now have a fair proportion of this dry land behind us.

Our belief in the power and importance of what some call "mere technical details" in understanding the overall functioning of complex systems cannot be expressed in a better way than through this comment of the physicist W. Pauli (reported by Gamow, 1966):

"This is to show the world that I can paint like Titian:



only technical details are missing."

V.3 Summary

Starting from a physical retina expressed as a square array of square receptors, the first stage of the processing proposed was a layer of TDU's providing the system with the transistence value of the position of every a.v.e. on the physical retina at every moment, whether or not it falls within the region of the physical retina currently being paid attention to. It was decided that the physical retina should undergo a constant tremor, one cycle involving a complete sweep over a (square) region of approximately 3/4 of a retinal field in "diameter". It was furthermore decided that the period of one tremor cycle should be adopted as processing moment, making sure that the light array's shift relative to the retina caused by its tremor would not lead to the detection of "false" transistence values by TDU's. The tremor was introduced to facilitate the line detection process taking place a little further on in the system. Apart from being given access to a layer of TDU's the physical retina's a.v.e.'s were also and more importantly given access to an attentional retina, the access to this attentional retina being selectively granted on the grounds of which region of the physical retina the particular a.v.e.'s come from. This meant that the system could choose to investigate all or some part of the

physical retina.

From the attentional retina it was decided that a.v.e.'s should be grouped into line segments before any attempt was made to compute useful transistence values for grouping purposes. A twisted pile was proposed to carry out the line segment detection task. The so-called twisted pile was designed to detect line segments very quickly by means of a kind of template matching process yielding as final result line segments specified by single cells in the twisted pile itself and characterized by these cells' respective positions and contents standing for positions, orientations, and sizes of detected line segments. The general-purpose character of this type of pile in everything that has to do with positions, orientations, and sizes was stressed, and the ease with which curved line segments could be included in the detection scheme was noted.

Having reached line segments with their respective positions, orientations, and sizes, it was decided that the time was ripe for computing transistence values as grouping criteria in deciding what goes into each of the three piles representing respectively the visual object, the visual sub-object, and the visual background. It was argued that quite a diversity of grouping criteria could

be obtained by placing TDU's in different ways in different piles, and we actually opted, as a first trial, for position-orientation specific TDU's in two different contexts: the context of the line segments' position-orientation relative to the retina (the "absolute" pile), and the context of the line segments' position-orientation relative to a chosen line segment (the "relative" pile). For separating out the visual object and the background, the system was provided with both an absolute and a relative transistence detection pile; and for identifying those line segments belonging to the object which should also belong to the sub-object the system was provided with a relative transistence detection pile only. The power and simplicity of transistence values as grouping criteria in cases of movement field effects was stressed. By allowing line segments to be selectively sent to the three piles on which the computation of actual motions would eventually be based we completed our account of the design of the first half of the system.

The first question tackled in our account of the design of the second half of the system was which M-characterizing features would best allow the visual object's movements (through a two-dimensional environment) to be detected. Nine such features were proposed for a start, three

requiring only consideration of the line segments making up the visual object itself and the other six requiring consideration of either the background's line segments (three features) or the sub-object's line segments (the other three) as well as the visual object's line segments. The first three M-characterizing features proposed were the visual object's global position, orientation, and size relative to the retina; the next three were the visual object's global position, orientation , and size relative to the background; and the last three were the sub-object's position, orientation, and size relative to the visual object. These last three features were introduced as "shape descriptors". Strategies for the detection and storage of each moment's value for every one of the nine different features were proposed and embodied in precise nerve net structures.

The next step was motion detection, and the general motion detection strategies proposed in Chapter IV were particularized for each one of the nine frozen features, nerve net structures being proposed for each motion to be detected.

We decided that our system should be allowed to eye-track objects involved in translatory movements relative to the organism. A sub-system was therefore proposed whose task

it is to take the visual object's detected translatory velocity (i.e. change of global position) relative to the retina and to transform it into a translatory velocity relative to the organism, this new velocity being used instead of the old one for further processing as well as for driving the eye-tracking itself. The sub-system was called the "Secondary System of visual motion detection", in contrast with the other parts of the system concerned with motion detection, these other parts being grouped under the label "Primary System of visual motion detection".

Translatory movement relative to the organism having replaced translatory movement relative to the retina within the system's set of nine detected velocities, our last move was to consider these velocities' speeds as nine new M-characterizing features and to provide each one with velocity detection facilities, thereby allowing accelerations and decelerations to be detected for each one of the nine "first level" motions.

As already hinted at when discussing the design of the line segment detection pile, we carried out a digital computer simulation of some aspects of the system described in this chapter. This simulation was meant to help clarifying the more complex parts of the system and

to check on their computational validity. Although the intrinsic incompatibility between existing digital computer software and nerve net systems such as ours made it impossible to carry out a complete simulation of our system within a reasonably short period of time we managed to implement a sufficient part of it to check at least on the computational validity of essential processes such as those carried out by the twisted piles and those responsible for running groupings and motion detection. In order to make this "preliminary" computer simulation possible we had to limit ourselves to a physical retina consisting of 16x16 receptive fields only. More details concerning which aspects of the system were simulated and how this was done can be found in Appendix C.

CHAPTER VI

Modelling biological systems

VI.0 Introduction

In this chapter we will discuss the relevance of the ideas which we used to design the system's micro-structure in the context of investigations on particular biological visual systems. The chapter will consist of two main sections, a first short one (Section VI.1) where the primitives used to design the actual micro-structure (i.e. TDU's, VDU's, and piles) will be considered, and a second (Section VI.2) where the actual micro-structure will be considered. The discussion of the micro-structure will be split in this chapter again into two main steps, the first (Section VI.2.1) dealing with the part of the system concerned with grouping a.v.e.'s into the single visual object and its two satellite v.e.'s (viz. the sub-object and the background), and the second (Section VI.2.2) dealing with the part of the system concerned with M-characterizing the visual object and computing motion.

VI.1 TDU's, VDU's, and piles

Firstly, and most superficially, there is the apparent physiological similarity between TDU's and the famous ON-OFF receptive fields found in the visual system of many animals by a number of investigators, starting with Hartline (1938,1940a, 1940b), together with the apparent anatomical similarity between piles and the highly organised mappings of receptive fields fourd mostly in the visual cortex of cats and monkeys (e.g. Hubel and Wiesel, 1962). However, while ON-OFF fields and cortical columns of receptive fields mappings are (partial) instances of TDU's and piles, the reverse is by no means true since TDU's and piles are concepts which attain a level of generality reaching far beyond concepts which are meant to account exclusively for "appearing or disappearing light in some given retinal position or even orientation" and "topographical correspondence of retinal positions in cortical regions". TDU's and piles are indeed concerned with totally general purpose issues where transistence is exhaustively and explicitely cared for (with its four possible values) for any value of any feature ("receptive fields" being only a <u>particular</u> domain for the computation of transistence) and where the organised mapping of values of features are expressed in conformity with general "featural topography" rather than being expressed in

conformity with particular "retinal receptive field topography" (in the particular case of receptive fields' positions on the retina, "featural topography" becomes retinal topography, but this is the only case where it is so).

What we want to emphasise is that in the context of physiological findings, TDU's and piles can be thought of as the two general classes of computational structures to which ON-OFF receptive fields and cortical column mappings of receptive fields respectively belong. Now the interesting point is that inter-related (general) classes of computational structures can be very helpful in analysing and relating particular computational structures such as those identified through physiological investigations. For instance the way in which TDU's, VDU's, and piles were brought together to form a motion detection unit could be the basis of an investigation aimed at breaking the apparent total independence and "self-contained" character of physiologically identified motion detectors. From the physiological evidence provided so far, motion detectors seem to be totally independent of any other identified detectors, such as ON-OFF detectors. As already argued in Chapter II it seems that the great weakness of the physiological understanding of visual systems is the lack of ideas about

how the different "specificities" identified relate and interact in their respective overall systems. The critical question therefore seems to be in what kind of context does such or such a specific detector lie and what does it have in common with other specific detectors. We believe that TDU's, VDU's, and piles could help to answer such questions.

In order to find relevant existing views on actual motion detection in biological visual systems at our level of interest we have to turn to Psychology, starting with Gestalt views on motion detection. We want to argue that the intuitive idea behind the "short-circuit" theory proposed by the Gestalt psychologists to account for motion detection in biological visual systems is not at all different from the intuitive idea behind our VDU or "travelling OFF" network. We believe that the main difference between the two models lies in the level of precision with which the basic intuitive ideas have been implemented, more than in the basic intuitive ideas themselves. The advantage of "travelling OFF" networks over "short-circuits" (or propagating electro-magnetic fields) is in the discreteness and explicitness of their functioning, together with their well-defined role in the larger context of running groupings and the even larger context of all visual groupings. It is however not clear

whether the "electro-magnetic propagation" spaces were thought of by the Gestalt psychologists as being merely "topographical" or if they were thought of as being "featural", the looseness of their model making both interpretations equally tenable. On the basis of the fact that they considered the whole of the optic sector as propagation space for their electro-magnetic fields, we can give them the benefit of the doubt concerning the "featural" nature of this space. However, no such concession can be made to Kolers' model of human visual motion detection since the "travelling" space of H-signals (i.e. those responsible for motion detection) is clearly a topographical space, where position on the retina is the only acceptable M-characterizing feature.

In short we can say that if Gestalt models and Kolers's model do account for some human visual abilities, VDU's account at least for these same abilities, and do it much more precisely and with a much wider applicability than these earlier models could.

It could be interesting to see in detail how VDU's can generate the types of human visual outputs on which the Gestalt models were made to stand. One type of visual phenomenon was obtained by presenting human subjects with successive exposures of two lights standing at some

distance one from the other, and by doing this in such a way that the subjects are made to perceive a single light moving back and forth; this phenomenon was called "apparent movement". In order to understand how VDU's can be made to see this "apparent movement" one only has to remember that VDU's are mechanisms designed to group OFF and ON values of features. When successive lights are exposed in different positions the successive OFF and ON values of position are fed into the VDU in the usual way, and an "apparent" translatory movement is detected. Less trivially, VDU's can also be made to account for the peculiar dependence of this apparent movement on relations between the time and space parameters of the stimulus structure as observed from experiments with human subjects. If one assumes a relatively slow information propagation rate (like that allowed by using the nervous influx as medium) instead of assuming instantaneity of propagation through VDU's (i.e. a speed of propagation which is high enough to make sure that the information gets to its destination well within the sampling moment) the system might be expected to display a peculiar dependence on space and time stimulus parameters. More precisely, if we take for instance a set of TDU's linked to some M-characterizing feature's pool of values and we give them a moment M (i.e. the TDU takes its input every M milliseconds), when a TDU's OFF output is sent through

the VDU's travelling network in search for an ON it has to pass through the whole net (i.e. through the whole set of values of the particular feature) within M milliseconds if M is to be kept as "moment" at the level of the VDU. Now the problem is that compared with a VDU there is very little "travelling distance" involved in a TDU, and if one wants to take the VDU as a basis for deciding on the size of the single moment one forces the TDU to work slower than it actually could. Thereby the TDU looses some of its power of temporal discrimination. One solution is obviously to have <u>different</u> moments for TDU's and VDU's, i.e. the temporal threshold for discriminating transistence values (e.g. ON-OFF sequences or "flicker") is set at a different value from the temporal threshold for discriminating velocities (i.e. grouping OFF and ON signals). The VDU's moment would of course be longer than the TDU's moment. This means that the OFF signal travelling through the VDU would be allowed to travel for a longer period than M and would thereby be allowed to meet ON signals that are detected one or more (TDU) moments after the OFF signal itself has been detected. Such a system would therefore be working <u>optimally</u> for stimulus conditions where the OFF and ON values of M-characterizing features are <u>temporally and spatially</u> set in accordance with the signals' speed of propagation through the respective "travelling OFF" networks. This

means that "the further away" the relevant values are "the longer" the period should be between the moment when one value is turned off and the moment when the other is turned on in order to get the ensuing OFF and ON signals to meet optimally, and the converse is also true. This direct inter-relationship of space and time parameters governing the setting of an adequate stimulus structure to obtain apparent movement is in fact the object of the first of the three classical laws governing the perception of optimal apparent movement by human subjects, laws which are attributed to Korte (1915) (e.g. in Forgus, 1966, p. 226). This first law says that for human subjects viewing successive exposures of two lights the condition for perceiving optimal apparent movement as time and space parameters are varied is that the distance between lights is varied in <u>direct proportion</u> with the time interval between the successive exposures (within certain limits of course). It might also be worth taking note of the fact that the flicker fusion frequency (or critical flicker fusion) for the human visual system implies a much shorter (TDU) moment than the (VDU) moment implied by the human system's temporal limits for apparent movement, the former being of the order of 20 ms while the latter varies between 5 and 10 times that much.

It seems straightforward enough to see that VDU's provide

at least as good a model as the Gestalt's short-circuit model, even in the case of the most "gestaltist" experimental context, that of apparent movement, but what is more interesting is that VDU's can lead to deeper issues. One of these can actually be set in this same Gestalt context and rests on the concept of "distance to be travelled by the OFF signal" in any given pool of value in order to trigger the expected apparent movement signal. This distance in the context of Korte's first law can hardly mean anything other than a "positional" or "metric" distance; on the other hand, when we talk of "distance" in the context of a VDU we refer of course to "featural distance" since the distance travelled by the OFF signal is expressed in terms of number of adjacent values passed by in the pool of values of any M-characterizing feature, the feature "position on the retina" being but a single particular case. Each different type of M-characterizing feature used by a visual system in its motion detection activities imposes a particular type of "distance" to which the stimulus structure yielding optimal (apparent) motion detection should conform, and "position" only represents one type of M-characterizing feature. The idea is therefore to extend the experimental scheme developed for translatory movement by the Gestalt psychologists to other types of movement. For instance considering the spatio-temporal conditions of discrete changes of stimulus

orientation leading to "apparent rotation" could lead to interesting generalisations of Korte's laws and could give some basis on which to assess the value of diversified global M-characterization as a valid aspect of a model of human motion detection processes.

As a final point concerning the suitability of our primitive computational concepts in the context of psychological investigations into the human visual system, we want to stress that, as was the case for physiological findings, the puzzling apparent diversity of the phenomena identified can be replaced by the much simpler common basis of our primitive concepts, and the easily diversified and classified uses to which they can be put. The labelling of psychological phenomena in the psychological literature to this day challenges the understanding of anyone interested in finding the slightest common thread in the field of motion perception. Some of the more "popular" labels are: real movement, apparent movement, induced movement, auto-kinetic movement, movement after-effects, movement field-effects, φ-movement, γ-movement (and many other greek-letter-movements), image-retina and eye-head movements, kinetic occlusion, translation, rotation, expansion, acceleration, pendulum movement, movement parallax, and the rest. It is hard to believe that each

such phenomenon has its private mechanisms; the human system has got to be more economical than this.

VI.2 The system's micro-structure

VI.2.1 From a.v.e.'s to the single visual object

Following the flow of information through our system we can first of all stress the similarity that seems to exist between the capacity of our system's attentional retina to bear on any portion of the physical retina and our own (human) visual capacity to attend to any part of the visual field. This could seem to be a rather trivial point but it is important since in the context of our system this attentional flexibility had to be introduced to allow for our "one visual object at a time" scheme to lead to a complete analysis of whatever scene happens to be presented. The point is that when the whole field of view is taken as object of attention the relative decrease in the precision of the description (as compared with the precision allowed when a very small area is attended to) is due to the fact that most of the time more line segments will have to be made part of the single visual object whose characterization is limited to a fixed set of global features, not to the fact that actual resolution at detecting line segments is lost. It is interesting to introspect on what happens when one tries, using one eye to start with, to attend to the whole visual field: there is a strange feeling of seeing everything and nothing at

the same time, and of being incapable of specifying individual objects although all their components (i.e. contours) seem perfectly clear.

A second similarity between our system and the human system is that "happenings" outside the current field of attention will be noticed by both systems. There is however a potential source of argument here in that the most widely accepted view about the human visual system's sensitivity to "peripheral happenings" is that it is a sensitivity to motion. We of course argue that transistence detection is all that is needed to attract the system's full attention or computational power, motion itself being detected when this full attention is applied to the retinal area concerned.

Moving a little deeper into our system we reach our decision of making the system group spatially adjacent a.v.e.'s into line segments before computing transistence as a basis for further grouping. The reasons motivating this decision can be easily recalled by having a quick look at Figures 19 and 20 again (pp.165,167). The similarity between our system's behaviour and that of the human visual system in this context is that when presented with the stimulus structure shown in Figure 19 (line A being shown at moment 1 and line A' at moment 2) the human

visual system seems to invariably see a single line moving (outcome shown in figure 19c), while when presented with the stimulus structure shown in Figure 20a (line A B C D E F being shown at moment 1 and line A' B' C' D' E' F' being shown at moment 2) the human visual system invariably splits the linear group of dots into two groups of three dots, one of which is seen as remaining still while the other one is seen as jumping over the still one. These results were however obtained through rather loose pilot studies using very few subjects and have to be considered as such. In the set of pilot studies that were carried out in this context it was noticed that although stimulus structures such as the ones shown in Figure 20a invariably yielded the result already mentioned it was very easy to alter the stimulus structure in such a way as to generate the other possible interpretation of the event, that of seeing a whole line of dots shifting to the right. This can be achieved by simply introducing a short "blank" period between the presentation of the two sets of dots. The effect is very convincing. What seems to happen in such a case, if we use our system as a model, is that the blank period causes OFF signals to be triggered by the TDU's associated with all the dots in every moment's stimulus pattern, these OFF signals being a sufficient basis on which to group all the dots together, and since the same type of grouping can be done on the basis of ON

signals when the second set of dots appears a single group of dots can be identified in each moment and seen to shift from left to right by the motion detection part of the system. This motion of the single whole cannot be seen by our system because it requires a system where OFF signals are allowed to run after ON signals for more than one strict processing moment, but it could easily be seen by a system where the use of nervous influx as information medium forces the motion detection moment to be longer than the transistence detection moment.

The requirement to derive line segments before doing anything else brings us to our line segment detection scheme and to a most interesting similarity between our system and several biological visual systems, including the human one. Among others, the human eye exhibits tremor. Not very surprisingly the human eye's tremor rate is reported to be higher than the flicker fusion frequency for human vision, and more interestingly its amplitude is very small indeed, being of the order of a couple of retinal receptors. Yarbus (1967) provides a few observations:

> "Of all forms of eye movements, tremor is the most difficult to study. The amplitude of the tremor is very low and its frequency very high.... Most records have shown that the amplitude of the tremor (its angular dimension) is comparable with the angular dimensions of the eye receptors, while its frequency varies from

> 30 to 90 cycles.... Analysis of the records I obtained of tremor yielded the following results. The amplitude of the tremor is 20-40 seconds of angle (1.0-1.5 diameter of the cones in the fovea). The tremor is composed mainly of movements whose frequency is 70-90 oscillations per second (much higher than the critical frequency of flicker fusion)". (pp. 113-115)

What we propose here is to extend the striking behavioural similarities between our system's eye tremor and the human system's eye tremor to functional similarities. In other words we hypothesize that human eye tremor is functionally meant to facilitate a template matching scheme for line detection. The most widely accepted view about eye tremor is that it ensures that the optical image projected on the retina at every moment does not stay too long on the same sets of receptors, a stabilised image on the retina causing perception to rapidly fragment and fade away. We do not see why this should require many different types of eye movements (e.g. slow drift, micro-saccades, tremor), and we believe that our hypothesis concerning the role of tremor in the human visual system not only fits perfectly well in the context of what is presently known about eye movements and their effects but also offers a basis for drawing a (new) distinction between the functional significance of tremor and of other eye movements in the visual system.

Apart from introducing tremor as a means of solving the

resolution problem encountered in designing the line segments detection pile, receptive fields were also introduced: the existence of receptive fields in biological visual systems is one of the best documented facts of Neurophysiology. It seems essential to stress here that fields were introduced in our system for strictly computational reasons, these reasons having nothing whatsoever to do with a concern for simulating biological systems by trivially including directly in our system whatever physiological findings are available, and this is why the presence of fields in our system is interesting in the present context.

After the line segment detection pile, we reach our system's transistence detection scheme providing running criteria for the ensuing grouping of line segments into background, object and sub-object. The claim in this context is that transistence values, computed on the values of appropriate features, are sufficient to lead to the adequate grouping of line segments "behaving" together through time. The hypothesis here is that the human visual system also uses transistence values as criteria for achieving such grouping.

That the human visual system does carry out groupings on running grounds has long been known to psychologists and

has in more recent years been demonstrated very clearly, as pointed out in Chapter II (e.g. pp.87,88). What can be gained from considering our system as a hypothetical model of human visual processing does not lie so much in the fact that it carries out groupings on running grounds as in the actual way in which it does it. Our hypothesis that groupings on running grounds ought to be achieved on the basis of transistence values challenges the widely spread view that (movement) field effects are actually achieved by the human visual system on the basis of locally detected velocities. While critically testing these two different hypotheses would have led us too far away from our main interests, we indulged in a very simple experiment aimed at showing that there are running field effects (or groupings on the basis of running criteria) which are experienced by the human visual system and which have to be dealt with through transistence detection for lack of anything else to be detected. The problem with field effects involving actual movements is that it is difficult to be sure whether the human system uses motion detection or transistence detection as source of grouping criteria, both being equally possible sources. However, if it is not possible to have situations which involve motion without involving local change (i.e. the object of transistence detection) it is perfectly possible to have situations which involve local changes without involving

motion. By showing that the human visual system does carry out groupings in such situations we can be sure that in some cases at least this system uses transistence values as grouping criteria, its ability to use velocities as grouping criteria remaining uncertain. The experiment

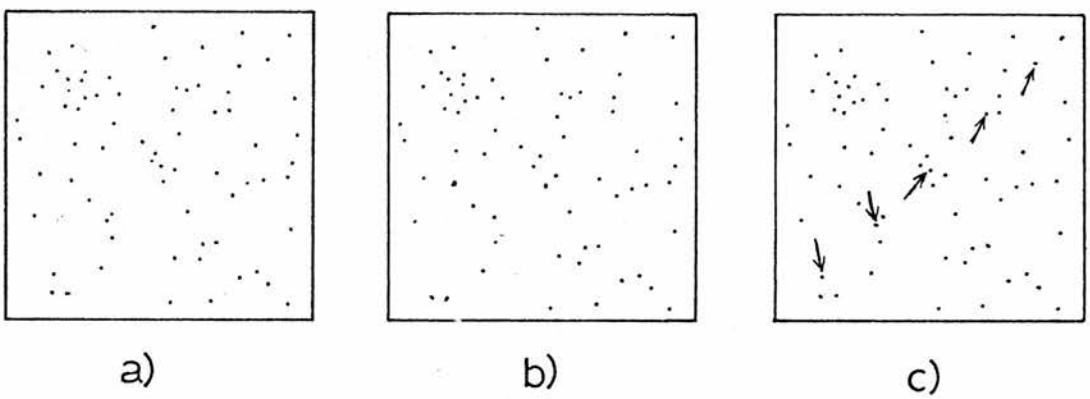


FIGURE 30. Random dot patterns used for demonstrating that the human visual system does use transistence values as grouping criteria.

designed to show just that involved watching a CRT screen where a set of randomly positioned dots were displayed and where at one moment a given sub-set of these dots formed of equally spaced lined up dots was made to go OFF, all the dots in the sub-set disappearing at the same time. The overall dot pattern was arranged so that no particular arrangements of bright dots or dark spots could be preferably noticed (on the sole basis of their existence) either before or after the disappearance of the sub-set of

dots. Figure 30a shows what the whole dispay looked like before the disappearance of the sub-set of dots, Figure 30b showing what it looked like after the chosen dots were turned off. In Figure 30c the whole initial display is shown once more but this time the dots which have been chosen to disappear (i.e. these which are not shown in Figure 30b) are pointed out to the reader by clearly marked arrows.

The question is of course whether or not a human observer will perceive what the group of disappeared dots looked like, i.e. whether or not the human observer will be able to tell, after the chosen dots have disappeared, that there was "a straight alignment of dots crossing the whole screen in a 45 degrees orientation", and that this set of dots just disappeared. In this experimental situation again very few subjects were tested, but they unanimously reacted in the predicted way: the length, orientation, and position of the disappearing group of dots were most of the time reported very acurately. The same experiment, but involving turning on the chosen dots instead of turning them off, was conducted just as successsfully. On the basis of the results obtained in such an experiment our point is that since actual motion is nothing but a succession of appearances and disappearances of the local elements making up the moving entity why would the human

visual system make its task harder by discarding transistence in favour of velocity when grouping moving elements in a scene? We do not see any reason why the system should do so, unless there are types of movement field effects to be detected where transistence, because of its very local nature, falls short of fulfilling the computational requirements underlying the task. However, the fact is that we have not come across (as yet) field effects which humans can experience which cannot be tackled by some transistence detection scheme. There is a lot of work left to be done in this context, the variety and range of complexity of running field effects with which human or animal visual systems can cope having received very little attention to this day. We want to suggest that this lack of attention is due to the almost total absence of criteria to express variety and complexity, and we want to stress that our system is a potential source of such criteria.

VI.2.2 M-characterization and motion detection

The first issue involves our ideas about shape detection. Three aspects of these ideas seem well-defined enough to act as precise hypotheses about human visual shape analysis, and these are firstly the fact that our system's single visual object is described in terms of three single global features (viz. position, orientation, and size) acting as reference for the description of shape, secondly the fact that shape itself is described in terms of local-global relationships within the visual object, and thirdly the fact that a sequence of local-global relationships is required to construct the complete shape of any non-trivial visual object.

Concerning the single position, the single orientation, and the single size which act as reference for shape description in the system at any moment the point of interest is the observed fact that the human visual system also seems to be limited in its choice of references. The most obvious case is that of the choice of axis of symmetry in situations involving many equally satisfactory such axes, or orientations. Attneave (1968) very interestingly discusses the human perception of triangles. The point made by Attneave (1968) comes out very clearly in the following excerpt from Attneave (1971):

> While planning an experiment on perceptual grouping I drew a number of equilateral triangles. After looking at these for a time I noticed that they kept changing their orientation, sometimes pointing one way, sometimes another and sometimes a third way. The basis for this tristable ambiguity seems to be that the perceptual system can represent symmetry about only one axis at a time, even though an equilateral triangle is objectively symmetrical about three axes.

Apart from the fact that the human visual system shifts from one interpretation to another, the interesting point in Attneave's observation is that one single global orientation is always taken as reference, even though in principle many different orientations allow for equally valid interpretations. This is of course exactly what our system does, one global orientation being all that is required for computing rotations.

The case of triangles is a special one because the three different orientations yielding symmetric interpretations also yield identical interpretations of shape (i.e. triangle). In the case of squares for instance all orientations yielding symmetrical representations do not yield similar interpretations, that is for a system which takes global orientation into account when detecting shape. This brings us to the second relevant aspect of our system's shape detection scheme, i.e. the local-global character of shape identification. Given the square showed in Figure 31a the vertical orientation

yields a symmetric interpretation, but so does the 45 degree orientation; similarly both the vertical and the 45 degree orientations of the "diamond shape" shown in Figure 31b yield symmetric interpretations.

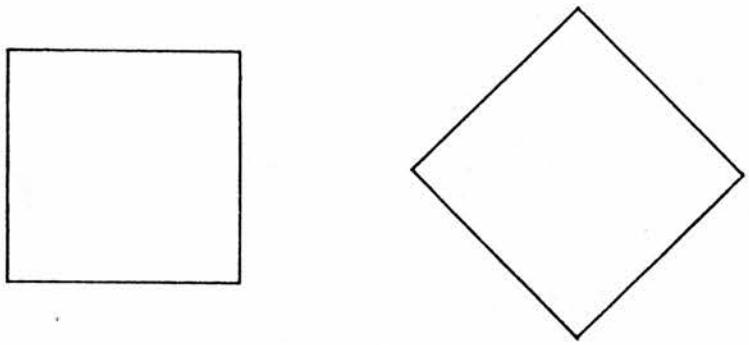


FIGURE 31. Square versus diamond shape.

If a shape description scheme involves only the detection of local-local relationships both shapes shown vertically in Figure 31 are described identically: they both involve four line segments of equal length linked to each other with the same ninety degree angles. However, in our system, the local-global relationships imposed on shape description will make the vertical interpretation of the two shapes completely different: Figure 31a will be described as a set of four line segments orthogonally set relative to the global axes (or global orientation), i.e.

a square, while Figure 31b will be described as a set of four line segments tilted at 45 degrees relative to the global axes, i.e. a diamond shape. The reader can judge for himself if Figures 31a and 31b are perceived as being similar or different shapes (in the vertical orientation). It could be worth while pointing out here that the human visual system seems to be very poor at evaluating local-local relationships within any visual scene; a major mistake of beginners at the difficult art of sketching consists in trying to follow the succession of local bits of the contour lines of whatever they are sketching, and this usually ends up by the "artist" being forced to start juxtaposing, half way through the drawing, elements of the scene which are in reality quite a distance apart, or vice-versa. Figuring out local-global proportions (and relationships in general) when drawing any non-trivial scene seems to be much more important and fruitful than figuring out local-local proportions. Coming back to our "square versus diamond shape" problem we realise that since shape description, in the human visual system as well as in our system, depends so heavily on global orientation we have to face up to the problem of deciding how global orientation is assigned. Our general solution to this problem was to say that the global orientation of any set of line segments would be the orientation which allows the most economic representation for the line

segments. In the partial computer simulation of our system (see Appendix C), orthogonality and symmetry were considered economic, and verticality was adopted to resolve ambiguities in cases where no previously adopted orientation was available to resolve them. These criteria seem to fit perfectly human visual responses to such stimuli as squares and diamond shapes, verticality being for instance the winning orientation in the isolated cases of Figures 31a and 31b but the diamond shape appearance of Figure 31a being easily kept intact as this figure is set into a smooth rotation which even goes through the "vertical square state" (the previous moment orientation being in this case predominant). But there is another even more instructive characteristic of global orientation selection by the human visual system, and this is the unquestionable influence of global orientation on local orientation (i.e. background on object, and object on sub-object). This influence has been noticed by Attneave again (1971) and is demonstrated in Figure 32 (after Attneave, 1971). The fact that the individual elements belong to sets which are symmetric along one particular axis is obviously determinant in the human visual system's choice of a global orientation for the elements of these sets.

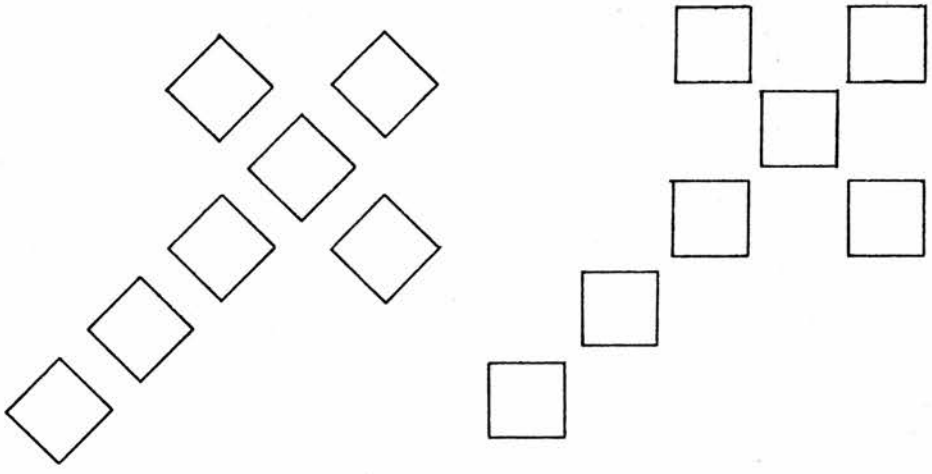


FIGURE 32. Influence of global lay-out on the choice of an orientation axis determining the perceived "shape" of local lay-outs.

As far as our system is concerned not only could the background and the object provide global orientations which are favoured by the object and the sub-object respectively in their search for a global orientation, but since an essential characteristic of our system is to analyse scenes by shifting different parts and sub-parts of it from one level to the other in the background/object/sub-object hierarchy one of the best hypotheses to go by when shifting levels is that the global orientation of the former background, object, or sub-object, is the same for the newly defined background,

object, or sub-object (i.e. the former orientations are favoured in cases of ambiguity). For instance if at the start our system considers the whole of Figure 32a as visual object it is bound to give it a 45 degree (clockwise) orientation, because of the very strong symmetry that this orientation allows. If at the next moment the system decides to choose some part of the whole figure as visual object, one set of four connected line segments for instance, the favoured orientation should surely be that of the former (larger) visual object, and since this orientation is "clockwise 45 degrees" the ambiguity in the new visual object between 0 degree (diamond shape interpretation) and 45 degrees (square shape) is resolved in favour of the 45 degree interpretation. This strategy can be applied in one given moment between the current background, visual object, and sub-object as well as it can be applied through successive moments between succeeding and inter-changing backgrounds, objects, and sub-objects. Our system is therefore perfectly at ease with the reported phenomena experienced by human subjects looking at shapes such as those shown in Figure 32.

This way which our system has of describing sets of line segments through sequences of descriptions in terms of different kinds of inter-relationships between certain

sub-sets of the line segments brings us to the third relevant aspect of our ideas about shape representation in the context of human visual processing. The issue revolves around ambiguous figures (such as triangles, squares, diamond shapes, and others) and takes its roots in the observed fact that human subjects do experience periodic shifts in their perception of shape when looking steadily at ambiguous figures. The classic example is the Necker cube, shown in Figure 33a, where two strikingly different perspectives of the cube are alternately perceived, only one perspective being seen at a time, and where the periodicity of the shift in perspective often appears to increase as one keeps looking at the cube.

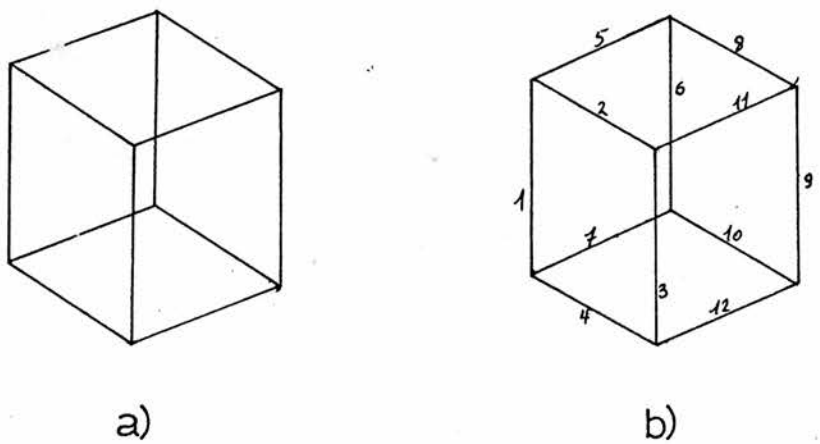


FIGURE 33. Necker cube.

Attneave's interpretation of this phenomenon (1971) assumes parallel detection of the two equally valid interpretations, the current percept alternating between the two co-existing interpretations. He indeed discusses the phenomenon in the following way:

> The most likely is that alternative aspects of the figure are represented by activity in different neural structures, and that when one such structure becomes "fatigued", or satiated or adapted, it gives way to another that is fresher and more excitable. Several investigators have noted that a reversing figure alternates more rapidly the longer it is looked at, presumably because both alternative neural structures build up some kind of fatigue. In some respects the neural structures behave like a multistable electronic circuit. (p.70)

And Attneave goes on to present an actual electronic circuit which models his explanation of the phenomenon.

We would tend to account for the periodic shift in the interpretation of ambiguous figures such as the Necker cube in quite a different way. We propose that the reason for the shift in interpretation is informational rather than energetic. The model proposed by Attneave is "energetic" because it calls for concepts (such as fatigue) which have to do with the system's power to carry (or process) information rather than with the system's way of carrying (or processing) information; this characteristic of Attneave's model is rather obvious from the analogy which he draws with an electronic circuit

where the main features of his model are expressed through electric charges leaking away or being built up. The alternative explanation which we propose is informational in that it rests on our system's strategies of visual analysis and is mostly related to the sequential nature of our system's shape analysis strategies as well as to the fact that our system gives only one object interpretation at a time. The point is that the different possible interpretations of an ambiguous figure do not co-exist in our system as they do in Attneave's model: they are generated one after the other as the system carries out its analysis of the different possible relationships between the different groups and sub-groups of line segments in the visual scene. The system's current interpretation depends on which group (or sub-group) of line segments have currently been chosen as background, object, or sub-object respectively. Referring to Figure 33b for instance if the system first chooses to group together (as a sub-object or as an object) line segments 1,2,3,and 4, the most natural interpretation, since these line segments form a perfectly valid non-occluded region, is to consider the side of the cube which they specify as standing in front of the other sides; this gives one interpretation of the ambiguous figure. But once one interpretation has been reached the system just does not stop there, it tries another one, many different journeys

being possible through any given set of line segments. If at the next moment line segments 3,11,9, and 12 are grouped together at the start, and are therefore considered as standing in front of the line segments making up the other sides of the cube, then the same interpretation as before will be reached. But when the moment comes where line segments 1,5,6,and 7 are grouped first and considered to be in front of the rest of them a new interpretation is reached, and the cube "reverses". There is no "fatigue" in the system, there is only a new absolutely valid interpretation yielded by the current moment's analysis. Furthermore the observed increase in the periodicity of the shift in interpretation as one keeps on looking at the figure can very well be accounted for by the fact that the system has got sufficient information to realise which journeys through sets and sub-sets of line segments are redundant (i.e. yield the same interpretation) as well as which ones lead to different valid interpretations, and can start avoiding the former and concentrating on the latter. Here again no fatigue, just optimal information processing.

There is another aspect of "reversing" figures with which our system is perfectly compatible. This has to do with the fact that changes in the observer's <u>point of visual fixation</u> can cause reversals to occur. This can be simply

interpreted by saying that since certain parts of the figure observed support one interpretation more strongly than alternative ones, when the observer "fixates" a given part of the figure this part's favoured perspective is seen. This fits perfectly with our system's shape analysis strategies as described above in the case of the Necker cube for instance. However, it was shown that reversals can occur in the absence of eye movements, and this has generally been taken to mean that something other than "the priority of the first part of the figure taken into account" should be included in the explanation of ambiguous figures' reversals. Attneave (1971) indeed argues:

> As Necker pointed out, changing the point of visual fixation may cause perspective to reverse. In the instances when the input is being matched against more than one schema visual fixation on a feature that is more critical to one representation than to the other may lock perception into only one aspect of the ambiguous figure. Since the percepts can alternate without a change in the point of fixation, however, some additional explanation is needed. (p. 70)

And Attneave goes on to propose the already discussed "fatigue" model. Our point is that the fact that reversals can occur in the absence of eye movements only means that changes in the physical retina's point of fixation are not determinant in the system's choice of perspective, and that what is determinant in the choice can either be the attentional retina's point of fixation

or even more precisely the actual grouping strategies (background, object, and sub-object) applied to what happens to be on the attentional retina. What we hypothesize is that in those cases where changes in the point of (physical) fixation cause reversals to occur the physical shift is accompanied by an attentional shift, and in those cases where reversals occur without changes in the point of (physical) fixation the system has only made an attentional shift. This allows us to have a single explanation for all aspects of the reversal phenomenon, this explanation being based on giving the priority to the interpretation favoured by the first part of the figure taken into account.

Moving on to the second main relevant aspect of our system in the context of modelling biological visual systems we leave the frozen domain of shape analysis to enter the running domain of motion detection. More precisely we turn towards our system's way of dealing with the visual object's movements relative to the background. The human visual system has long been known to psychologists to be able to detect the motion of an object relative to its surroundings even though this object is perfectly stationary relative to the system's retina itself. Some of the purest demonstrations of the human ability to perceive this type of relative motion are none other than

Duncker's classical "induced motion" experiments discussed in Section II.3 (cf. pp.81,82). What was found interesting in Duncker's experiments in Section II.3 was the fact that these experiments showed that quite apart from using "position relative to the retina" as one of its M-characterizing features the human visual system also used "position relative to the natural frame of reference" (or "position relative to the background" in our terminology). Our interest in Duncker's experiments in the present context is however quite different and actually lies in the evidence which these experiments provide regarding what factors are determinant in deciding what is moving relative to what in cases of relative movement between visual entities. The main conclusions reached by Duncker in this context were drawn from many experiments; the following excerpt from a translation of Duncker's paper (1929) describes the two most significant experiments and also presents Duncker's general conclusions.

> The setup was as follows. In a dark room two projection lanterns cast spots of light upon mirrors which in turn reflected the light from behind upon a screen.....One (spot of light) was held steady and the other moved horizontally....(a just liminal motion was used).
>
> It was found (1) that in this case the total phenomenal motion of both objects equalled their phenomenal displacement; (2) that four of the six observers saw the <u>fixated</u> point as moving regardless of whether this point was the

> objectively moving one or not. (For the other two observers the motion of both points was more symmetrical.) (3) That when no fixation instructions were given, the experienced motion was usually symmetrical.
>
> To test our earlier hypothesis regarding the role of "localization" the following experiment was carried out.
>
> One of the objects was a point.....and the other a contour rectangle.....The point was inside the rectangle. The speed of objective motion (whether of the point or of the rectangle) was always the same. In all cases the unmoving object was the one fixated. There were ten observers.
>
> In all cases.....the strongest movement was that observed in the point, <u>not</u> in the rectangle. Ordinarily (e.g. with two points) it is the fixated object which is seen to move; here the fixated rectangle never moved. It is clear that the influence exerted by fixation in our earlier cases is not sufficient to overcome the figural factor of enclosure. The issue now is, not which object is <u>fixated</u>, but which one is <u>enclosed</u> by the other. Thus we have encountered two determinants of perceived motion :(1) other things being equal, there is a greater tendency for motion to appear in the <u>fixated</u> object than in the non-fixated one; (2) the same as regards the enclosed rather than the enclosing object. (pp.164-165)

This whole discussion should be very reminiscent of Attneave's discussion of the role of the fixation point in reversing ambiguous figures (cf. p.297). Here again indeed fixation is argued to be a determinant in the visual analysis concerned, but a determinant which is left behind in favour of other ones because of its apparent failure to account for some experimental outcomes. Duncker argues that "topological enclosure" is one of these more powerful determinants. What we want to propose here

is that fixation was found by Duncker to be insufficient to account for all experimental outcomes because it was limited to physical fixation, and that if one considers instead attentional fixation as the determinant factor deciding what moves relative to what, topological enclosure slips back to the status of a secondary determinant and all experimental results can be accounted for through a single primary determinant. This might sound like quite a drastic statement, especially since Duncker's induced motion is taken nowadays as a classic demonstration of the importance of spatial relationships (such as enclosing-enclosed) in human vision. It is important to realise here that we do not discard topological determinants altogether in our claim: we merely hypothesize that they would have a secondary role, being for instance involved in deciding what the attentional retina should concentrate on or what should be taken as visual object, but by no means necessarily imposing enclosed objects as visual objects for instance. The important point is that it is what happens to be chosen by our system as visual object which is seen as moving relative to the rest of the scene projected onto the attentional retina, and this whether or not topological enclosure (or any other determinant for that matter) has been used in the grouping process. If this aspect of our system is a good model of the corresponding

part of the human visual system, this one should be able to overcome topological enclosure through appropriate attentional fixation. Duncker reported that when the contour rectangle was still and the enclosed dot was moving no observer fixating the contour rectangle ever reported seeing the rectangle move: they always saw the dot move. Our belief is that even though these observers did fixate (physically) some point on the contour rectangle, their attention was on the enclosed dot, and naturally, according to our hypothesis, the dot was seen as moving. We believe that if the observers had steadily attended to the contour rectangle as well as physically fixated it they would have seen it move. We did try the experiment and it worked very convincingly, although a lot more concentration was needed to get the contour rectangle move when only the enclosed dot was physically moving than vice-versa. This is probably due to the fact that the contour rectangle is harder to control as visual ooject. Whatever the reason may be the important point remains that topological enclosure cannot strictly be claimed to be a primary determinant in the context of relative motions. Less importantly but still quite interestingly our hypothesis accounts much more precisely than Duncker's for the results which he obtained in his experiment (reported above) with the two single spots of light. According to our interpretation his second finding (2)

simply shows that four subjects out of six had the fixated dot under attentional as well as physical fixation while the other two subjects were physically fixating one dot but their attentional fixation was on both dots; similarly, concerning his third finding (3), the absence of (physical) fixation instructions obviously made most subjects consider the two dots as a single visual object.

A last point of interest concerning our system's way of handling "induced" motion as compared with the human way is that our system explicitly computes the visual object's motions relative to the background through three different M-characterizing features. Although in the above paragraphs we have been referring mainly to relative translations (i.e. changes in relative position), "induced" motions perceived by our system include relative rotations (i.e. changes in relative orientations) as well as relative expansions and contractions (i.e. changes in relative size). That the human visual system actually detects such a variety of induced motions is easily demonstrated, even though very little evidence concerning this problem seems to be available from the psychological literature. As far as relative rotations are concerned the only evidence which we found in the literature is provided by Duncker himself in a short section of the paper already referred to, and as far as relative

expansions or contractions are concerned we did not find anything in the literature. In order to confirm Duncker's finding that relative rotations are experienced by human observers we looked for situations involving such rotations and found a few very convincing situations in movie films, and in order to make sure that relative expansions and contractions are available to human vision we made an animated cartoon involving such motions and found that human observers reacted very convincingly in the expected way.

Let us now turn to the third and final main relevant aspect of our system to be discussed in the context of modelling human visual processing, and this time it is the whole of the Secondary System of visual motion detection which is concerned.

When we were designing the first version of the secondary system, it occurred to us that there existed an experimental setup for accurately controlling the critical stimulus parameters underlying our system's way of handling motion. As the system developed, we found that this setup offered enough diversity to enable us to check on almost every one of its abilities. Every time a new ability was given to the system, we tried to design a way in which we could use the experimental setup to check it.

The aim of this exercise was obviously not to find an experimental scheme for studying our own system: what we had in mind was to use the experimental setup as a complete paradigm for the investigation of the human secondary system, to check the validity of our system as a model of it.

The basic phenomenon behind the whole experimental paradigm is as follows.

In our system, once the eyetracking has been "initialised", the secondary system only needs null-velocity inputs from the primary system in order to continue to track and see an object in uniform motion (see section V.2.2). A situation that satisfies only these critical needs would be a situation where, once some tracking has been "initialised", the tracked object is physically motionless relative to the organism as well as relative to the retina. If this situation could be found we would have a case of apparent motion where in fact a stationary object is being eyetracked as well as seen in motion.

The idea which we used to obtain such a setup is the following one. If our system's eye is made to track a target moving along under a stationary row of identical

and equidistant objects, and if this row of objects is illuminated stroboscopically at a flash rate which is similar to the target's rate of "passing under" the successive objects in the row, then we have a situation where the objects of the row are motionless for the primary system (i.e. relative to the retina). The reason for this is quite simply that, since every flash of light only occurs when the eyes have moved one inter-object distance along the row (i.e. the same set of retinal "spots" always receives the row of objects) and since all objects in the row are identical, there is no retinal clue left to infer displacement.

In such a situation our system will "see" the elements of the row moving along with its eye and the initial tracking target, and if it decides to transfer eyetracking from this initial target to one of the elements of the row it will be able to carry on tracking this element and observe its motion until the very end of the row.

We therefore carried out an experiment where we used a row of 250 equidistant black dots as our row of objects, and a small spot of light running over the row of dots as our tracking target (see Figure 34). We used 30 minutes of arc as interdot distance, 50 Hz as stroboscopic flash rate, and consequently 25 degrees per second as the speed

of our tracking target.

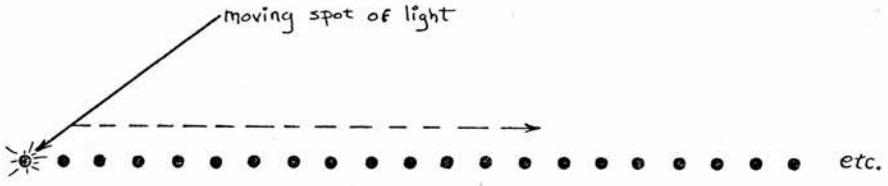


FIGURE 34. The basic stimulus setup.

We observed that as soon as our eyetracking was "initialised" (using the moving spot of light) the whole row of black dots jumped into motion. We then concentrated on the black dot which the spot of light seemed to be riding and we took the spot of light away: the tracking continued smoothly, and the black dot appeared to be moving as before. We were then well "locked" in a tracking loop which was taking us happily and "unconsciously" from one stationary dot to the next. With a little paractice we found out that we could trigger "the phenomenon" extremely easily, using the moving spot for a fraction of a second only, and track to the very end of the row. Although the phenomenon was extremely

convincing by itself, we nevertheless recorded (by EOG) a few eye movements to make sure that eyetracking was actually taking place after the withdrawal of the initial tracking target. The results clearly showed that eyetracking was indeed taking place. Hereafter "the phenomenon" will refer to the sustained tracking and the consequently perceived motion of the elements of the row once the initial tracking has been removed.

The phenomenon can be experienced under quite a wide range of frequencies of stroboscopic lighting. For example, in the case of our first experimental setup where the interelement distance was 30 minutes of arc, the phenomenon could be obtained under flash rates chosen anywhere between 10 Hz and 150 Hz.

Theoretically the upper limit of the range of "permissible" flash rates (given an interelement distance) depends directly and only on the physical eyetracking mechanism's speed limit: if the flash rate (combined with the given interelement distance) requires a critical tracking speed which exceeds the power of the tracking system, then clearly the phenomenon cannot be obtained. To avoid a common misunderstanding it might be worth stressing the fact that the flicker fusion frequency of the visual system involved is not a critical factor in

setting the upper limit of the range of "permissible" flash rates. One might indeed fall into the natural trap of equating continuous "perceptual" lighting and continuous "physical" lighting: we therefore stress that the phenomenon is produced on the basis of a discontinuous physical lighting only and that consequently the elements of the row do not have to be perceived as discontinuously lit to allow it. There is therefore no surprise in realising that the phenomenon can be obtained easily with flash rates well above fusion.

On the other hand, the theoretical lower limit of the range of "permissible" flash rates depends quite a lot on "perceived" discontinuity. Here the matter is a little tricky, but the basic idea is the following one. As the flash rate is brought below fusion (i.e. below perceptual continuity) and reaches a point where the elements of the row do not trigger "still" signals anymore, the phenomenon is still allowed for some lower frequencies since the secondary system accepts ON signals as meaning "immobility" (see Section V.2.2); but as still lower flash rates are selected the ON signals are pushed further and further towards the edge of the sampling period until the stage is reached where the time interval between flashes is so long that a complete sampling period is deprived of its ON signal, thereby breaking the continuity of tracking

by dropping below the lowest possible "permissable" flash rate.

The fact that the phenomenon could be obtained at flash rates much below fusion therefore constitutes an experimental outcome in favour of the hypothesis that the secondary system works on ON signals as well as on STILL signals. This hypothesis is also supported by results obtained by Gregory (1958) in an experiment in which subjects were asked to track a physically moving self-luminous target in a room illuminated stroboscopically at very low frequency (5 Hz). Gregory's subjects reported an apparent motion of the room along with the tracked target. However, in such a setup, the apparent motion is continuously interrupted by the OFF phase of the lighting cycle, and although the objects in the room can periodically be seen as moving with the eyes, they nevertheless move away from the centre of the retina and are not replaced by exact copies of themselves in the retinal spot that they were occupying previously. For these reasons the setup allows neither the eyetracking nor the apparent motion in the absence of the self-luminous tracking target. Although, in our context, this experiment provides evidence which is more partial than the evidence offered by the phenomenon on the treatment of ON signals, it is nevertheless important.

We have seen above that a critical eyetracking speed (in degrees per second) for eliciting the phenomenon can be derived from multiplying interelement distance (in degrees) by flash rate (in Hz). We will now consider the critical speed worked out this way as the basic critical eyetracking speed, and we will discuss two ways of deriving from it new critical speeds (for a given interelement distance and a given flash rate).

First, any multiple of the basic critical speed is itself a critical speed (its actual "permissibility" depending on whetheror not it exceeds the power of the eyetracking mechanism). The reason for this can be grasped by realising that any speed of eyetracking which preserves the immobility of the elements of the row (relative to the retina) is a critical frequency, and that given a flash rate and an interelement distance any multiple of the basic critical tracking speed preserves this "immobility". Obviously the perceived speed of the apparant motion is expected to conform with each different critical tracking speed.

Secondly, submultiples of the basic critical speed are themselves critical speeds. The basic idea behind this theoretical expectation is here again that the submultiples of the critical speed preserve the immobility

of the elements of the row relative to the retina although they create singular side effects. Let us take for instance a situation where we have a row of black dots 30 minutes of arc apart on a white sheet of paper under a flash rate of 100 Hz, and where the initial tracking target is set at a speed of 25 degrees per second (i.e. half the basic critical eyetracking speed). As soon as the system "initialises" its eyetracking using the moving target, the following stimulus pattern falls on the retina. A first flash of light projects the row of dots onto the retina in those spots marked "F" (for first) in Figure 35. Since the eye has travelled only half of the interdot distance when the second flash of light comes, the row of dots is projected onto the retina in those spots marked "S" (for second). Now when the third flash of light comes, since the eye has travelled a complete interdot distance since the first flash (i.e. twice half an interval), the row of dots is projected again onto the retina in those spots marked "F", and when the fourth flash of light comes the row of dots is projected onto the retina in those spots marked "S", and so on. So what happens is that the same two sets of spots on the retina are successively and repeatedly exposed to the black dots. This creates a situation where, if the flash rate is such that two successive flashes happen within one sampling moment, our system's eye will experience the phenomenon on

one single (simultaneous) row of dots where the dots are twice as close and twice as numerous as they would be in the case of eyetracking the same physical setup at the basic critical speed. Also, since in this situation "every other flash" projects "whiteness" on the black dots' retinal spots, our system will find the black dots "sort of greyish".

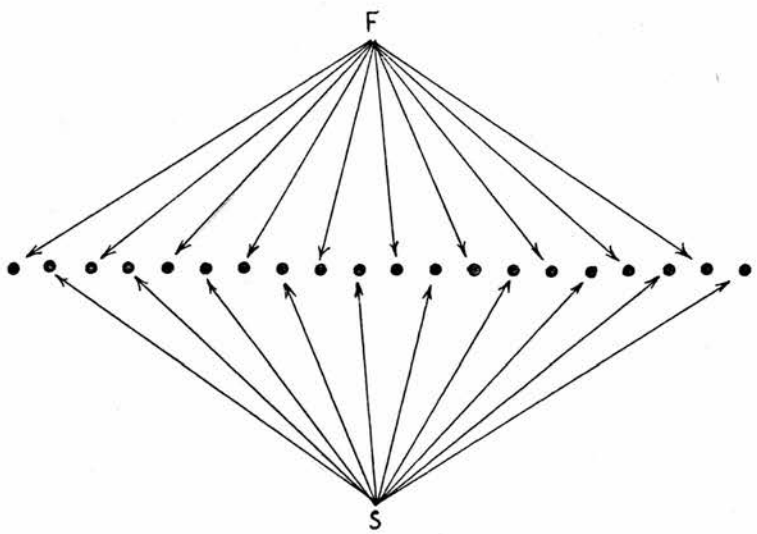


FIGURE 35. The "doubling" effect at half the basic critical eyetracking speed.

The general picture that emerges from this discussion is that when we divide the basic critical speed of eyetracking by some factor N (N being an integer) we get a critical eyetracking speed at which the system perceives a row of N times as many elements (with an interelement distance N times as short) as when it is tracking the same

physical setup at the basic critical speed.

In the present state of our theory the phenomenon should be obtained for any value of N as long as the "range limitations" discussed above are respected, and as long as the elements of the row receive enough light to allow a decent contrast. However, we feel somehow that the maximum value of N should otherwise be set at the point where the elements of the row get so close to each other that they form a perceptually continuous straight line (we want to allow the phenomenon for partially overlapping objects).

Now the following results were obtained from preliminary controls made on human subjects:

1-Multiples of basic critical eyetracking speeds were observed to be critical speeds themselves.

2-Submultiples were also observed to be critical speeds; the change in the number of elements of the row and the interelement distance as a function of N was obvious. The highest "N" which we managed to reach with the particular rows we were using was N=4.

A rather obvious point is that the discussion of how

critical tracking speeds can be worked out for a given flash rate and a given interelement distance is also valid (at least as far as the general idea of multiples and submultiples is concerned) for deriving how critical flashing rates can be worked out for a given tracking speed and a given interelement distance, as well as for deriving how critical interelement distances can be worked out for a given tracking speed and given flash rate. This is hardly surprising since these three parameters are so much related in the production of the phenomenon; but we felt that the fact was worth noting all the same.

Until now the phenomenon has allowed us to check on the ability to perform eyetracking in cases of uniform motion only. Let us now have a look at how the phenomenon can help us investigate the ability to perform eyetracking when nonuniform motion is involved.

Since nonuniform motion involves changes of velocity and since flash rate is one factor conrolling tracking velocity in the context of the phenomenon, a relevant question would be: how would our system react to a progressive change of flash rate once the phenomenon has been triggered on some row of equidistant elements? Such a change would in fact create a situation where the flashes of light would occur before (if the flash rate is

increased) or after (if the flash rate is decreased) the eye has travelled through a complete interelement distance. In this case, as opposed to the case of multiples or submultiples of the basic critical tracking speed (or flash rate), different spots on the retina are exposed to the elements of the row from moment to moment and therefore the elements' immobility is lost. This causes the primary system to detect changes of position relative to the retina and to compute a velocity from them. This velocity is then sent to the secondary system (in fact to the MSS) where it is combined vectorially with the current eyetracking velocity in order to work out the new velocity of the element relative to the organism, velocity which becomes the next eyetracking speed. This means that in the event of a progressive change (obviously within certain quantitative limits) of flash rate our system automatically accelerates or decelerates accordingly (depending on the type of change), and the phenomenon is kept going.

For similar reasons, given a constant flash rate, progressive changes of interelement distances will trigger an automatic acceleration or deceleration of the eyetracking speed, thereby keeping the phenomenon going (see Figure 36a).

Following the same type of argument we also reached the conclusion that the phenomenon will be kept going automatically in the event of a change of direction of eyetracking. In other words if the row of elements goes on straight for a while but then slants away at an angle of, for example, 20 degrees (see Figure 36b), the system's eye should be able to take the tracked element over the bend without causing the phenomenon to break.

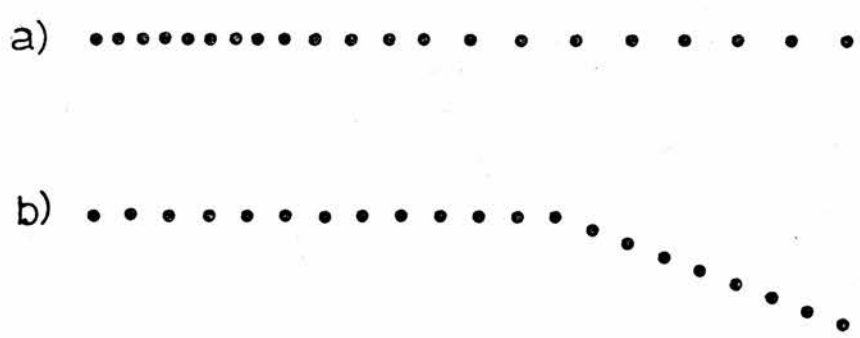


FIGURE 36. (a) Speed change stimulus; (b) direction change stimulus.

In those cases where changes of velocity are involved it becomes important to keep the attention of the visual system well focused on the tracked element, since, as we explained in Section V.2.2, the whole power of the system is required when changes in the tracked object's velocity occur.

The following results were obtained from preliminary controls made on human subjects:

1-Given a row of equidistant black dots, the phenomenon was kept going under many rates of progressive increase and decrease of the flash rate. Corresponding perceptual impressions of acceleration and deceleration were observed.

2-Given a constant flash rate, the phenomenon was kept going as the tracked dot was taken over increasing (tracking left-right) and decreasing (tracking right-left) interdot distance (see Figure 36a). Corresponding perceptual impressions of acceleration and deceleration were observed.

3-Given a "bent" row of equidistant black dots under a constant flash rate, the phenomenon was kept going as the tracked dot was taken over the bend (see Figure 36b).

4-The phenomenon broke down (i.e. subjects suddenly found themselves facing a set of stationary dots) in all three situations when the subject's attention was not focused on the tracked element when the changes came. This piece of data comes from reported

"impressions" only, since attention is a very hard thing to measure.

The fact that we can deal with changes of velocity has an interesting consequence as far as critical eyetracking speeds are concerned. It in fact means that there is no need to set the motion of the initial tracking target on the exact theoretical critical value: any approximation of this critical value which falls within the correction range of the secondary system will trigger the phenomenon (the corrections being automatically and immediately done by the secondary system). This makes the phenomenon much easier to obtain: for instance one only has to track one's own finger while moving it at different speeds below the row of elements and a sufficient approximation of some critical speed (the basic one or some multiple) will soon be found.

We have just seen that changes of position of the tracked element relative to the retina affect the human system very much like they affect our system. In our system, these changes are the only ones to be coped with by the secondary system; indeed, we saw in Section V.2.2 that the only changes (detected by the primary system) which the secondary system considers are changes in global position of the visual object relative to the retina. So this

means that no other type of velocity undergone by the tracked element will interfere with the phenomenon itself in the case of our system. Since in our system there are eight types of velocity other than the positional velocity relative to the retina, we would have a long way to go to discuss the problem thoroughly. We therefore decided to illustrate the general technique of discussion in this context by choosing two out of these eight remaining velocities: "rotational" velocity of the object relative to the retina (based on changes of orientation relative to the retina), and "positional" velocity of the object relative to some other object on the retina.

To deal with rotational velocity relative to the retina we will consider a row of equidistant elements consisting of a set of line segments spread out in a spatial succession reproducing the temporal succession of those orientationns involved in a rotational movement (see Figure 37a). Since the object to be tracked is a line segment, the visual system has to work out a (global) position for it: in the present case this position has to be the centre of the line -if the line is to be tracked. Once the phenomenon is triggered in such a situation, our system "sees" the tracked object as a rotating line involved in translatory motion. The experiment was carried out with human subjects and the hypothesis was confirmed.

Now to deal with positional velocity of the tracked object relative to another object on the retina, we will consider a row of equidistant black dots, where each dot is "framed" by a black circle which changes slightly its position relative to the dot from one dot to the next (see Figure 37b). The interesting outcome here is that, once the phenomenon has been triggered, our model "can see" the motion of the tracked dot relative to the circle even when the direction of eyetracking is the opposite of the direction of the "relative" motion, and it can carry on with the tracking all the same. Again this was confirmed in an experiment with human subjects.

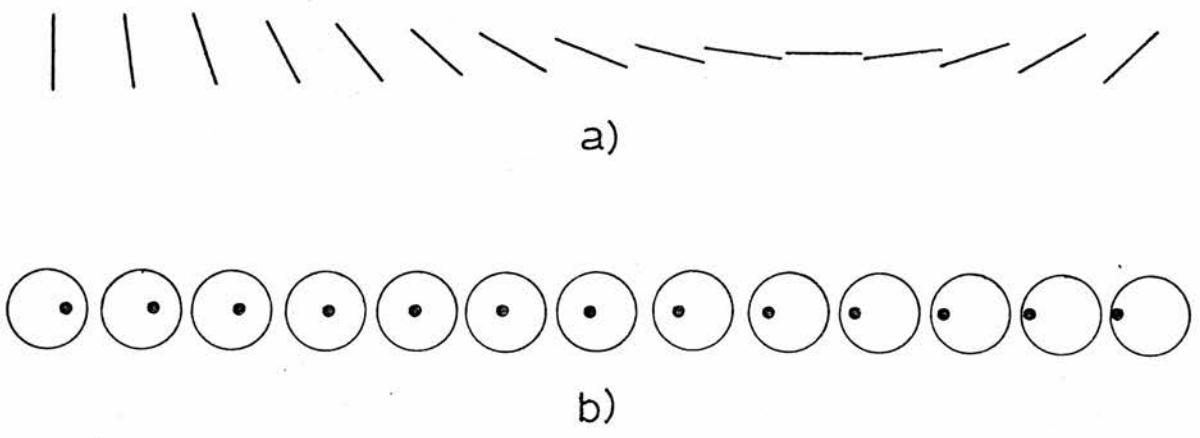


FIGURE 37. (a)Rotational motion stimulus;(b) relative translational motion stimulus.

Finally, we saw in Section V.2.2 how our system is able to

transfer its attention (i.e. some of its computational power) to other objects when the tracked object is in uniform motion. We also saw how this allowed the transfer of the actual eyetracking from the "old" object to the "new" one. This capacity was in fact required for actually allowing the phenomenon in all the situations we have dealt with so far: we indeed had to go <u>from</u> the initial tracking target <u>to</u> some element of the row in every case. However, the process is under much better control in the following situation.

Let us consider a "pile" of rows of dots where within each row the dots are equidistant, but where each row has got an interdot distance slightly shorter than the interdot distance of the row just below it(see Figure 38).

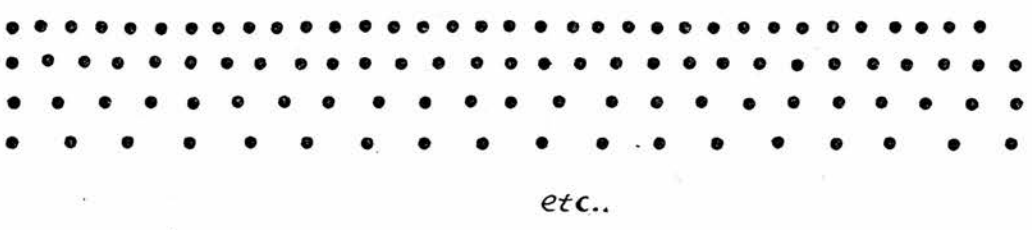


FIGURE 38. Eyetracking transfer stimulus.

If the phenomenon is triggered on the top row, our system's eye can then "jump" from row to row right down to the bottom row of the pile, and climb back up again without ever causing the phenomenon to break. This involves, within the system, quite a subtle exercise. However subtle the exercise, the human capacity to do the same was successfully controlled. After a little practice one could control the phenomenon beautifully, leaving the tracked element to the good care of the "autopilot" (the ASS in our model) and observing other elements (moving at different velocities in the rows nearby and at the same velocity in the local row), or transferring the tracking itself to one of those observed elements (changing row if the observed element is in another row, remaining in the same row if the observed element is in it). An interesting outcome of this type of ability is that given a single row of equidistant dots, one can jump back and forth in this row allowing the phenomenon to be kept up indefinitely. We made a few eye-movement recordings of such a case, and the recordings clearly showed that after intervening saccades are made the eyetracking is automatically carried on as before.

This discussion about our Secondary System in the context of modelling certain human visual abilities and about the ensuing experimental findings has been published

(Lamontagne, 1973). At the same time we came upon a paper by Stoper (1973) where related problems were discussed. Having engaged in correspondance with Stoper it was found that the basic illusion behind the standard case of the apparent movement phenomenon reported in our paper had also been noticed by Stoper and reported in his doctoral dissertation (Stoper, 1967, p. 147). Stoper however concentrates mainly on what we consider to be the "pre-phenomenon stage", i.e. the stage where the tracking target which is used to initialize the phenomenon is still being tracked by the observer. We indeed only start talking of the phenomenon once the observer has made the eye-tracking "jump" from the "triggering target" to one of the apparently moving objects. The fact that this can actually happen (i.e. that subjects can "ride" apparently moving objects in the absence of anything changing position relative to the environment in which they stand) is as far as we are concerned the most revealing fact and is what makes the whole situation quite a powerful experimental paradigm. Moreover we believe that most instances of the phenomenon discussed above lie outside the predictive range of Stoper's model, and some instances even contradict it. It indeed seems from Stoper's discussions that he believes that command velocities sent to the oculo-motor system are not vectorially combined with the corresponding detected <u>retinal</u> translatory

velocities (in cases where these retinal velocities are different from "0"). He indeed argues:

> 1- The pursuit movement. Most visual phenomena observed during the pursuit movement can be explained by assuming only the existence of the "oculomotor motion signal" described above. In general, the only effect the existence of this motion signal has is to cause the perceived motion of objects which have a stationary image on the retina." (p.170)

If, as Stoper seems to be inferring from his experiments, retinal velocities and eye-tracking velocities are never vectorially combined the "phenomenon" should break down as soon as a "combination" is required to carry it on, e.g. whenever the system has to cope with accelerations, decelerations, changes in direction, jumps to new apparently moving objects with different velocities. Since the phenomenon persists in these situations Stoper's position appears to be somewhat over-generalized. However our model also falls short of accounting for all observed aspects of human visual processing in eye-tracking situations. The vectorial analysis which we talk about in our paper does not seem to be carried out in all cases by the human visual system, but the boundary between when it is applied and when it is not applied remains somewhat unclear. A typical situation lying outside the range of our Secondary System's "vectorial analysis box" is for instance, as Stoper pointed out, the Filehne illusion (i.e. seeing the physically still background move in the

opposite direction when eye-tracking a physically moving object); a very convincing equivalent of the Filehne illusion can be created using a very simple apparent movement setup where two spots of light are made to move at the same (rather high) speed along orthogonal paths on a CRT screen and to cross each other in the centre of the screen; one is moving horizontally and the other one vertically: if an observer is made to eye-track the horizontal one while "putting" his attention on the other one he unmistakably sees the latter move along a diagonal path across the screen, and the impression is so strong that most subjects which we used could not believe that the dots were actually moving on orthogonal paths. This result clearly demonstrates that there are indeed situations where the vectorial analysis combining retinal velocities and eye-tracking velocities breaks down (at least at a "phenomenal" level). We can therefore conclude by saying that Stoper's findings and our own findings are complementary rather than contradictory, and that one task which remains to be done is to find a way of integrating the respective models.

A last remark about the Secondary System is that the phenomena described in the paper have in the last two years generated quite a lot of experimental investigation in different research centres. Some results from these

investigations have in fact already been published and the reader is referred to Heywood (1973) and Korn (1974) for more details. Since these experimental investigations were triggered by the results of our attempt at designing a working visual motion detection system we take them as evidence of the fruitfulness of this kind of approach.

CONCLUSION

As stated in the introduction, the purpose of the work reported in this dissertation was aimed at:

1-creating computational concepts which cover the widest possible range of particular visual motion detection systems;

2-using these concepts to articulate a complete particular working system;

3-looking for abilities of this particular system to serve as computational model of visual motion detection in particular biological systems.

Concerning our first preoccupation, although our theoretical effort has not led to anything like a comprehensive theory of visual motion detection it has yielded sufficiently general concepts to make it possible for us to

1-delimit the territory of visual motion detection in the land of vision as a whole,

2-state the more precise problems of motion detection as such in a way which opened up a pool of well defined different possible solutions, as well as offering a single explicit context within which to compare and assess previous efforts at solving issues related to our problem,

3-express particular solutions clearly, right down to the smallest computational details required to make them operational in terms of working visual motion detection systems.

Concerning the second, although a lot of work remains to be done before we can claim to have a working visual motion detection system which can compete with the human system a particular system has been designed where progress towards this goal has been made at two different levels:

1-at the level of the system's macro-structure where we solved general problems concerning when, where, and how motion detection abilities should be set to work in a complete visual system;

2-at the level of the system's micro-structure where we solved problems concerning the detailed mechanisms underlying the motion detection abilities themselves.

The system's macro-structure has been designed under assumptions of (a) monocularity, (b) discrete sampling of the light array in both time and space, and (c) bright line drawings on dark background as only legal stimuli. The system's micro-structure was constrained further by being designed to tackle two-dimensional motion only.

Finally concerning our third preoccupation, the most interesting results were obtained by treating certain aspects of the system as models for human visual information processing abilities, and these results ranged from sharpening already existing views on such abilities (e.g. raising the ideas behind the Gestalt's short-circuit theory to new levels of precision), going through proposing new interpretations of available data (e.g. proposing that human eye-tremor is functionally "meant" to facilitate a kind of template-matching process), and carrying right up to using new interpretations to predict new visual phenomena (e.g. predicting the existence of a new type of apparent movement situation providing an extensive experimental paradigm for studying certain aspects of human visual eye-tracking abilities and related phenomenal experiences).

However, if we have solved a few problems in the course of our investigation we have also left many of them unsolved. These problems, occupying very precise places in our framework, offer obvious targets for future investigations. Grouping them into global projects, and listing these projects from lower to higher level ones, we can say that we intend to

1-strengthen the "shape analysis" aspect of the system,

2-simulate the whole system on a digital computer using a much larger (simulated) physical retina than the one assumed for the preliminary simulation already carried out (see Appendix C),

3-design nerve net structures which can handle a developmental process by which the two-dimensional motion detection micro-system proposed in the dissertation can become a three-dimensional motion detection micro-structure, and investigate further the possible developmental role of "running analysis" in organising "frozen analysis" (see Appendix B),

4-design and implement modes of higher level intervention in the lower level processes which our motion detection system consists of (which could be expressed in more fashionable terms by saying that we intend to design and implement the top-down counter-part to our bottom-up motion detection system), and

5-test the flexibility of our general-purpose conceptual framework (or language) by trying to design alternative micro- and macro-systems dealing with motion perception, and use the results of these tests to widen the scope of our computational theory of visual motion detection.

Adding to this the concern for modelling biological visual processes, which should be kept in mind as one proceeds to

tackle the above mentioned issues, our plans for further investigation on the basis of the dissertation can be considered more or less complete. All on our own it is obvious that we could not pursue these plans very far, but we believe that some enthusiastic collaboration could lead surprisingly deep into them. The collaboration which seems to be needed involves ideally five fields of research: Artificial Intelligence, Computer science, Electronics, Physiology and Psychology. We believe our problem to be sufficiently well-defined to allow fruitful multi-disciplinary team-work to be done. The role of Physiology and Psychology in such team-work is obvious enough, many precise hypotheses to be checked experimentally being already available from the explicit achievements reported in this dissertation. As far as Electronics is concerned its interest lies mainly in the context of the implementation of proposed nerve net structures, electronic hardware allowing claims on speed of processing to be tested and allowing real-time simulations to be carried out. The design of electronic hardware is in fact the ultimate test of the proposed nerve net structures although it cannot really be undertaken safely before software simulations have shown that the logic of the proposed nerve nets is correct and leads to the desired results. Our twisted pile would for instance be a good project for hardware construction, a

computer simulation of this nerve net structure having already been successfully carried out. The role of Computer science lies of course in the context of the software implementation of proposed nerve net structures. Finally and most importantly the role of Artificial Intelligence consists in actually coining adequate computational concepts and designing on their basis the particular nerve net structures through which the desired computations should be carried out.

APPENDICES

APPENDIX A

Discrete sampling strategies

Basically there seems to be three possible options regarding which type of discrete sampling of the light array falling on its retina could be used by a visual system: the option of using a "gap" sampling strategy, which leaves some "territory" unsampled since the chosen discrete pieces of space and/or time do not cover the entire space and/or time dimensions (within the limits of the system of course); the option of using an "adjacent" sampling strategy, which uses discrete pieces of space and/or time covering the entire space and/or time dimensions, each piece covering an exclusively "private" territory; and finally the option of using an "overlapping" sampling strategy, which uses discrete pieces of space and/or time covering space and/or time dimensions in their entirety, but with each piece covering some part of the "territory" of some other piece(s).

These three options are best understood when discussed in the context of one dimension alone: let us consider the temporal dimension, keeping in mind that most of what will be said about time bears on space as well and can be used to shape the spatial as well as the temporal sampling problem in one's mind.

Although in some respects it seems far from being obvious, the slicing up of time is as necessary to visual analysis as is the slicing up of space. In fact our whole system rests as heavily on timing as on "spacing", since the "moment" specifies an a.v.e. as importantly as the "position" does. The idea of defining time slices, or moments, has to do with the need to decide on criteria for establishing simultaneity of events, everything falling within a time slice being considered simultaneous. Time slices are therefore units of resolution in the same sense as retinal positions (or space "slices") are units of resolution. In the absence of time slices our visual system could never decide what it is presented with, and the spatial sampling would carry on to eternity, everything being simultaneous.

We have already separated sampling strategies into three main types: the "gap" sampling, the "adjacent" sampling, and the "overlapping" sampling. In the temporal dimension the "gap" sampling can be compared to a cine-camera type of sampling, where the retinal information is gathered for some period of time after which a "shutter" closes blocking off the light from the retina for a new period of time before the retina is again exposed to light. This is

the simplest type of strategy, but the temporal "gap" in the sampling introduces a restriction, viz. the potential information in the light array during the "gap" is lost. This can give rise to illusions of the kind one experiences in movie films when seeing the classical "reverse rotation of coach wheels" effect. With the second type of sampling strategy, "adjacent" sampling, no gap is introduced between the successive sampling moments. In this case the moments "touch" each other, the second moment starting when the first one ends. Finally, "overlapping" sampling brings the time slices even closer together by making the moments encroach on each other's territory. If one considers the succession of such moments dynamically (as a single "moving" moment) instead of statically (as some pile of partly overlapping moments) one can describe this type of moment, along with Allport (1968), as being a "travelling" moment. Allport proposes the "travelling" moment as the type of sampling moment used by the human visual system. A visual representation of the different sampling strategies is given in Figure A-1.

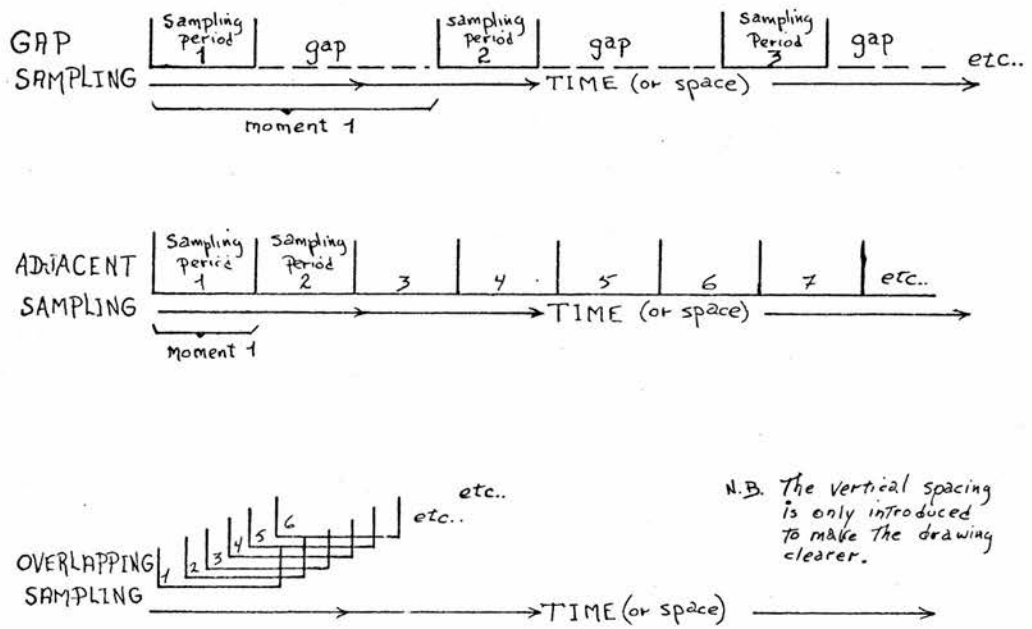


FIGURE A-1. The three main types of sampling strategies.

Now, the most powerful type of strategy is the use of a "travelling", or "overlapping", moment: the closer the beginnings of the moments follow each other, the better chance the system has not to miss anything happening in the observed world (for a given "moment duration" of course). But one has to avoid continuity; we have not chosen <u>discrete</u> moments only to revert to continuity by allowing them to overlap in all possible ways and amounts. The critical question is should we introduce an amount of time which <u>cannot be cut down</u> by any overlapping of moments? In other words, referring to Figure A-2, should we introduce a minimum time value X which can never be overlapped from moment to moment?

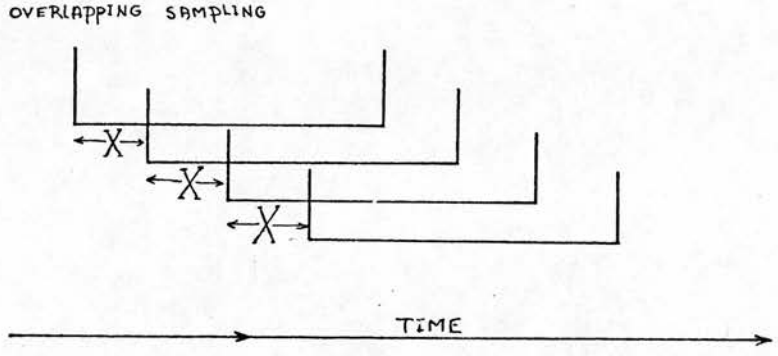


FIGURE A-2. Minimum time value (X) for non-overlapped period of overlapping sampling.

<u>If we do introduce</u> such a minimum value the discrete character of the sampling is saved, and in such a case the succession of moments can be better pictured in terms of a "jumping" moment rather than as a "travelling" one. Such a sampling strategy could be thought of as a two-level one, where adjacent moments of value X are used for a primary sampling and where larger overlapping moments are formed by grouping the primary moments at a secondary level of sampling, as shown in Figure A-3. This type of sampling accounts well for Allport's experimental results, and we therefore favour it as model of human visual sampling.

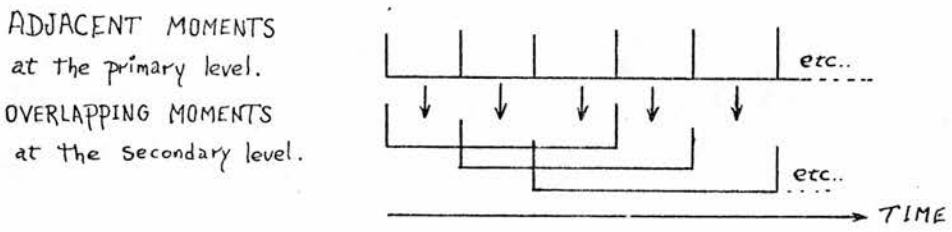


FIGURE A-3. Combined adjacent/overlapping sampling.

On the other hand if we do not introduce such a minimum value (X) we are talking about a continuous process, and then trouble starts. The kind of problems encountered in the land of continuity is best portrayed by Zeno's classical paradox of Achylles and the turtle. The paradox goes like this : Achylles is one mile behind the turtle, they are both travelling along the same path, in the same direction, and at constant speed, but Achylles is going twice as fast as the turtle. Even though Achylles is going faster than the turtle it can be argued that he will "never" catch up with it because whatever the distance still to be covered by Achylles to catch up with the turtle at some given moment, there exists a next moment where there will still be some distance to be covered by him to catch up with it (e.g. when Achylles has covered the mile that first separated him from the turtle, it is still half a mile ahead; and when he has covered this half mile the turtle is still one quarter of a mile ahead, and so on to infinity). The argument of course rests on the ability of the observer (part played by Xeno in his paradox) to "sample" the respective positions of Achylles and of the turtle at temporal intervals which can be reduced to infinity. It is indeed the case that a visual system which uses a first sampling moment of four minutes (say that Achylles runs at the constant speed of one mile in four minutes) and cuts its moment in half every time a new sampling is done would not see Achylles overtake the turtle for as long as the moment can be cut in half. If the moment can be cut in half to infinity, then it is true that this system would never see Achylles catch up with the turtle. But if, like all other physical systems, this system is physically limited to the evaluation of a minimum amount of time, i.e. a period of time that cannot be divided, when this moment is reached the paradox breaks down and Achylles is finally allowed to overtake its "momentarily saved" opponent. The point is that each

sampling moment requires a computational effort from the system, and this takes time. If the time spent on analysing successive moments is compatible with the time taken by the observed event to unroll itself, then everything is fine; but the shorter the moment, the more numerous moments are within a given period of time, and the faster the system has to process them if it wants to keep up with the event. In the limiting case, that of Xeno's argument, since an infinity of moments have to be processed in a given period of time, the system has to process them at an infinite speed. Physical constraints on all known information processing media make such a system practically inconceivable. Furthermore, the "diminishing" sampling moment implied in the statement of the paradox is far from being computationally economical since the evaluation of speeds in such a context requires non-linear operations to be carried out (i.e. the duration of any given moment, since it changes constantly, cannot act as unit of time over which distances travelled can be directly used to indicate speeds). In other words not only does it seem impossible to actually design a system where the sampling moment could be reduced to infinity, but it also seems to be a bad policy to allow the sampling moment to be changed at all within a given motion detection task.

Now among all discrete temporal sampling strategies using an unchanging moment, if it is true that "jumping overlapping" sampling is the most powerful it is also true that it is the most complex to handle, because of the embedded cycles which it implies. The "gap" sampling strategy, on the other hand, might be weaker but is certainly much easier to use. On realising that the gap sampling's drawbacks were very secondary to our purpose we decided to start designing our system on its basis. In the spatial domain the "adjacent" sampling strategy turned out to be the easiest and most adequate one at least as a starting point, and we adopted it in the first part of our work, transforming it into a kind of "combined adjacent/overlapping" sampling when we introduced receptive fields in the second part of our investigation.

APPENDIX B

Motion detection and visual development

In our attempt to bring into light the exact inter-relationships that can or should exist between running and frozen groupings in visual analysis we came to realise that in a developmental (or evolutionary) context running groupings can advantageously be made to rule the relationship.

The question is how can a visual system equipped with the simplest schemes of analysis acquire more sophisticated ones. In other words, how can a system decide "for itself" what the grouping criteria and the characterizing features to reach higher levels of representation should be? In our case the simplest schemes of analysis are those which yield a.v.e.'s. The starting point could therefore be a system which stops its analysis when a.v.e.'s have been obtained. What can we do with such a system? The v.e.'s which this system knows are totally frozen, and, in a way, have no real spatial dimension either since they are "point-like". However, since two of their features offer grounds for comparison and grouping, namely position and moment, there is hope for creating new generations of v.e.'s exploiting the potential multi-dimensionality of space and time. The system can either explore space alone, or time alone, or both "at the same time". Exploring space alone (i.e. limiting the exploration to a single moment) can hardly be done safely because time cannot be stopped, and exploring time alone (i.e. limiting the exploration to a single position) would be of little interest since time itself flows steadily in a single direction only (unlike "positions", which can vary in all sorts of interesting ways). The best option therefore seems to be to explore the two dimensions together, that is, to analyse the possible relations of a.v.e.'s which differ both in position and moment. Let us take for instance the case where we have nine successive retinal positions (adjacent and in a straight line) which are stimulated with light in nine successive moments. There is only one a.v.e. detected per moment, which is the simplest spatial stimulus possible. If we take the first two a.v.e.'s presented (i.e. the one at moment-0, and the one at moment-1), we have what we want, i.e. two a.v.e.'s which differ both in position and moment. By grouping them (under the criterion of temporal adjacency for instance) we are doing a running grouping, and any feature characterizing this group will be a running feature. By looking at the differences in position from moment-0 to moment-1 the system might then realise, given the appropriate data and process structures, that there

are two ways in which positions can relate: a quantitative one, the distance between the two positions, and a more qualitative one, the direction in which the second position appears relative to the first one. These features of the v.e. consisting of the two a.v.e.'s with different positions and successive moments can then be made standard to any successive a.v.e.'s in time, and they can be called respectively "speed" and "direction" of translatory motion. Now the important point is that those running features are not saying something which is only relevant in the "running" dimension : they in fact contain a lot of potentially very interesting information concerning the possible elaboration of features in the purely "frozen" domain. It is indeed the case that the spatial relations discovered through time could also be made to exist within but a single moment. The only thing one has to do is to FREEZE the running features. The frozen equivalents of direction and speed of translatory motion are ORIENTATION and SIZE of a line segment. The interesting thing which is happening here is that from a "zero-dimensional" object (i.e. a "point-like" a.v.e.) made to run in a one-dimensional space (i.e. along a line) we have obtained, by "freezing" the motion, a one-dimensional object (a line) with characterizing features (orientation and size) which can themselves be set into the running mode and allow motion to occur now in a two-dimensional space. It is indeed the case that orientation and size are both multi-valued features, and thereby allow motion to be "hooked" on them; the result is a v.e. which can "rotate" and "expand or contract" as well as "translate". These running features again lend themselves to "freezing" processes which can bring the system's analysis to the level of two-dimensional objects with characterizing features which ... but we have now reached the limit set by our two-dimensional retina: the two-dimensional multi-valued features characterizing our two-dimensional object cannot possibly be made to move in a third dimension as such and thereby allow three-dimensional objects to come out of yet another straightforward freezing process. So does this mean that three-dimensional (hereafter 3-D) motion and 3-D objects cannot be reached in the context of a system such as the one which we have just started to build? Not quite; there seems in fact to be a way in which the system's development can still be pushed into three-dimensionality, and where running features are crucial as ever. The idea is of course to start from the two-dimensional (hereafter 2-D) motions of the 2-D objects already obtained and to coin frozen features which will bear on the 3-D aspect of these objects. Since our visual system at this point knows nothing about 3-D, a system which knows about it will have to be used in combination with the 2-D visual system, and this system can easily be a motor system. So

let us see in which way our 2-D visual system combined with a 3-D motor system could be made to reach 3-D vision.

The idea is to give a 3-D interpretation to 2-D changes by using the motor system's ability to actively change the input structure by either moving the organism containing the eye or moving objects sitting in front of the eye. Whenever the motor system is about to affect the 3-D structure of what the eye is looking at it tells the visual system about what is going to happen in 3-D (this information is available from the motor command itself) so that the 2-D changes which are going to happen as a result of this motor intervention can be used as indications of 3-D changes. For instance, if the motor system manipulates an arm and hand holding some object so that the object is made to go away from the eye, the command sent to the hand is something like "move away (along the z-axis to the retinal plane) from the eye", and the 2-D change which will occur and be detected by the visual system, namely a change in size of the object, can be used as an indication of motion in depth. Then by bringing back to a frozen status the feature "change of size", the system obtains a frozen feature, "size", which can be made to bear on a 3-D characteristic of the object, namely its position on the z-axis. The important point to realise here is that this cannot be achieved without motion; the 3-D description of the environment will be made to emerge completely from running evidence. Frozen information alone seems hopeless in designing such a developmental scheme; if one thinks for instance of the possibility of having the hand holding the object steady at a certain distance from the eye, one realises very quickly that even though the 2-D visual system can be informed of the fact that the object lies at such a distance from the eye there is no way in which "size" can be identified as being the critical 2-D feature in establishing "position" in the third dimension: a change has to be introduced to specify it. It is of no use either to have "someone else" moving objects in front of the eye; it is indeed obvious in such a system that the motor system altering the stimulus structure has to be part of the same organism which is meant to interpret 2-D change in 3-D terms, otherwise the crucial information about 3-D itself is not made available to the visual system.

A very interesting possible outcome of this interaction between motor activity and 2-D change or motion detection is to be found in the case where the organism itself is made to move in its environment. What happens then is that with a little exploration the system can rapidly realise that objects which are furthest away undergo, relative to it, translatory movements which are slower than those undergone by objects which are closer. The

system can even realise that "distance away" is directly linked to "speed of translatory movement" (in those cases where objects in the environment are "still" relative to this environment itself, of course). Now the interesting point about this discovery of "motion parallax" is that it provides all the necessary information to justify an investment in binocularity. It is indeed the case that having two eyes processing at the same time two slightly different perspectives of a scene is the frozen equivalent of having one eye processing them through related successive moments (i.e. in terms of running features). In other words 3-D representation through binocularity could also be achieved through this same basic idea of "freezing on the basis of running evidence".

The motion detection system which we developed seems to offer quite an adequate background for applying these ideas about visual development. First, the versatile nature of piles offers ideal grounds on which to "run the frozen" and "freeze the running". Secondly, the general-purpose primitives used in the system to achieve the detection of different types of movement can facilitate greatly the process of raising the system's level of motion detection. Let us consider the case of going from two-dimensional to three-dimensional motion detection on the basis of an interaction between two-dimensional motion detection and three-dimensional motor activity. As argued above, the interaction yields the two-dimensional visual (multi-valued) features on which motion in the third dimension is to be computed, but says nothing about how the motion itself should be computed. If the system does not have any "idea" about general characteristics of motion detection, quite a lot of work remains to be done before motion in the third dimension can be computed. But if it uses general-purpose primitives for the motions which it already detects, the system (implicitly) knows how motion is computed. So the only thing which remains to be done once the appropriate M-characterizing features have been identified is to set these features into appropriate piles already equipped with transistence and velocity detection schemes.

APPENDIX C

Computer simulation

C.0 Introduction

Simulating a nerve net system on a digital computer can lead to two very different types of problems. The first type involves genuine issues about computational faults or underspecifications inherent to the nerve nets themselves, and this is really the type of problems which we are interested in finding out and solving. The second type of problem involves issues which are totally irrelevant to the nerve net computations themselves, these issues having to do with language incompatibilities making certain nerve net computations or structures extremely hard to express in computer language. These computer representation problems of course have to be solved if one wants to reach the genuine issues concerning nerve net computations themselves, but one should always bear in mind throughout the following discussion that the restrictions imposed on the simulated system in order to solve computer representation problems by no means apply to the actual nerve net structures being simulated.

C.1 TDU's, VDU's, and piles

Concerning the general-purpose concepts of TDU's, VDU's, and piles the following straightforward computer programming schemes were used. Firstly, all data structures were arrays accessed in terms of traditional coordinates (as many coordinates specifying any cell in an array as there are dimensions in the array). All values of features on which transistence was to be computed were therefore found in particular specific cells in given arrays, each value being specified by the actual set of coordinates determining its cell's position in the array. The existence state at any moment of the value represented by each cell was available from the content of each cell at any moment. TDU's were implemented where required in such arrays by providing each standard cell firstly with an associated cell, the "previous moment" cell, where the standard cell's existence state at the previous moment was saved, and secondly with a program which used straightforward conditional statements to decide at any moment, on the basis of the pair of existence states obtained from the "standard" cell on the one hand and the "previous moment" cell on the other hand, if the transistence value of the given feature's value represented by the standard cell is either an OFF, or an ON, or a STILL THERE, or a STILL NOT THERE. Such a setup

can be used for computing transistence on any standard cell, i.e. any value of any feature, and in those cases where transistence has to be computed for all values of multi-valued features, since these values occupy whole dimensions of arrays (i.e. sub-arrays) one can think of the associated "previous moment" cells and associated TDU routines as associated arrays instead of as associated cell-units.

Multi-valued features' associated TDU layers can of course act as "travelling OFF networks" for velocity detection. The precise way in which VDU's were implemented is that OFF signals were made to search for ON signals along pathways expressed in terms of (discrete) sequences of coordinates, any particular such sequence representing a particular direction of motion, and the number of cells passed through from the beginning to the end of the journey leading to an ON signal representing the speed of motion. The only problem with this computerised version of our VDU, besides the obvious fact that parallel processing is not allowed, was that in these cases where the desired sequences of values to be "passed through" could not be expressed in terms of sequences of coordinates as easily as they were expressed in terms of nerve networks the computer version of the processing was relatively slow going. However, this does not affect the validity of the ideas implemented, any more than it affects the expected speeds of processing which the nerve nets themselves would allow, were they hard-wired.

In short we can say

1-that TDU's were implemented by using straightforward conditional statements,

2-that piles were implemented through traditional multi-dimensional arrays accessible in terms of coordinates, and

3-that VDU's were implemented by expressing travelling OFF networks in terms of sequences of coordinates defining pathways through the associated TDU layers of the piles concerned.

C.2 From a.v.e.'s to the single visual object

The first computer representation problem was that of deciding what the size of the simulated physical retina should be. The problem is that since the computer works serially the more cells this retina contains the longer it will take to do whatever has to be done with them. It was in fact found out that if we wanted to have results within reasonable computing time we could hardly go for a retina containing more than 16X16 positions or receptive fields. Having found that 4X4 retinal cells was a reasonable size

for a receptive field we decided to use as simulated physical retina a (two-dimensional) square array of 64X64 square cells. It is important to realise how small such a retina actually is (the human retina having approximately 137 million receptors (Yarbus, 1967)). However, the few line segments which could be projected on this tiny retina allowed the system to work on stimulus structures which were rich enough for us to check our main ideas, especially since we were more concerned with running than with frozen diversity.

We started by defining a physical retina consisting of a two-dimensional array of 64X64 bit-cells (i.e. cells which can either contain a 1 or a 0). The stimulus line-drawings (entered by means of a "touch screen" device) were recorded on the retina by setting to 1 every cell on which they happened to fall, all other cells remaining in the 0 mode. Since by definition the different positions on the retina are defined in terms of fields and not in terms of cells the standard stimulus line had the width of a receptive field, i.e. four retinal cells.

Figure C-1 shows the retinal mapping of a typical input set of line segments (5 line segments in this case); those cells containing diagonal lines should be understood as being set to 1, and the empty cells as being set to 0.

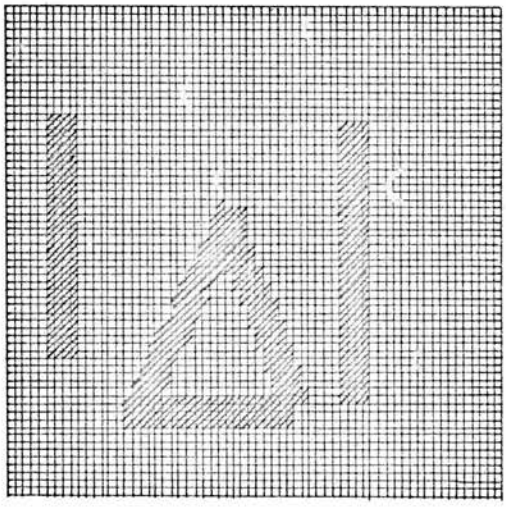


FIGURE C-1. Typical input array.

The way in which <u>tremor</u> was implemented is by recording on the physical retina the overall mapping of the stimulus line segments as the retina is shifted eight times (plus

one starting position) by discrete amounts of one retinal cell, following the pattern shown in Figure C-2. For

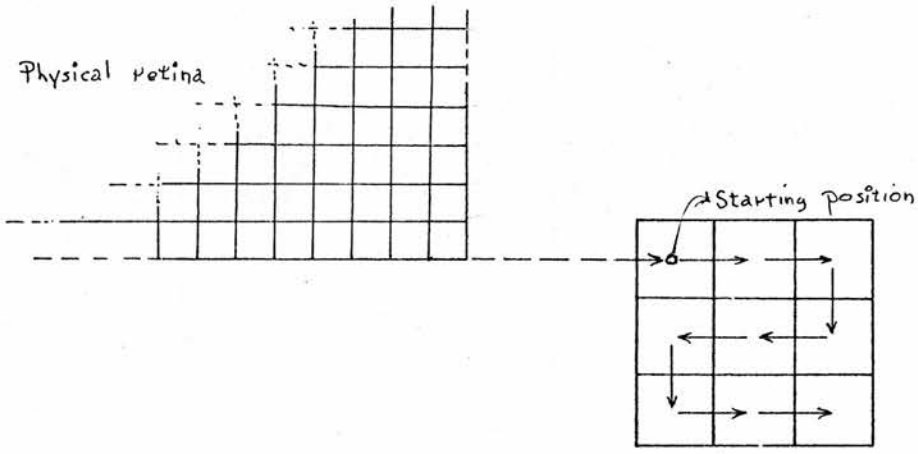


FIGURE C-2. Tremor pattern.

any set of input line segments there were nine mappings on the physical retina before its pattern of noughts and ones was considered complete for any one processing moment. For instance adding the extra mapping caused by tremor to the single mapping of the stimulus line segments shown in Figure C-1 would yield the complete retinal array of noughts and ones shown in Figure C-3.

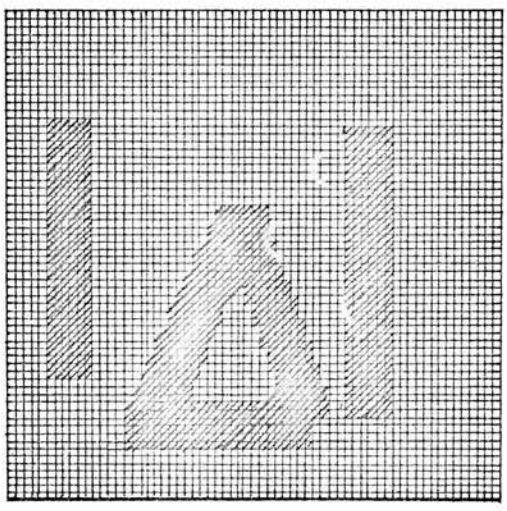


FIGURE C-3. Input array over one complete tremor cycle.

So the physical retina was "exposed" to the stimulus for one tremor cycle and then "hidden" for the equivalent of

two tremor cycles, or sampling moments, before being exposed to some stimulus again, all retinal cells having been reset to 0 again.

Now the result of each moment's sampling, in terms of an array of noughts and ones, such as the one shown in Figure C-3, was sent on the one hand to the array of TDU's from which warnings of anything happening on the physical retina outside the field of attention were issued, and on the other hand to the attentional retina according to some selection strategy. The layer of TDU's mentioned here was implemented in the way already described but was not used in the simulation because we always took the whole of the physical retina as object of attention. The reason is of course that our physical retina was already so small that any further restrictions of retinal size in going from the physical to the attentional retina would not allow for enough diversity to be of any interest. So for all practical purposes in the simulation we considered the physical and the attentional retinas to be equivalent; from now on we will refer to the input array of noughts and ones simply as the retina.

The next step was the implementation of the line segment detection pile, and this is where we met the second and most important computer representation problem. It has to do with the fact that a twisted three-dimensional array does not have a simple equivalent in present-day computer terms. In our nerve net pile the linearity of processing (e.g. mapping of retinal cells, vertical inhibition) and hence the speed of processing were ensured by the twisted structure of the pile itself; in the computer there is no such thing as a "twisted array", and our twisted pile had to be mimicked through non-linear operations defining the desired twisted path through a straightforward (non-twisted) pile. This allowed us to check on the computational validity of the nerve net twisted pile, but it tended to be a rather time consuming simulation, even more so since all operations had to be carried out serially.

Concerning the dimensions of the computerised pile we used an "orientational" resolution of five degrees and since our retina had already been designed to allow for 16X16 fields only the simulated pile consisted of 36 layers of 16X16 cells each. The line segment detection scheme within this pile was simulated as follows. The retinal tremor cycle involved nine "shift positions" for every sampling moment, and for each such "shift position" the content of each retinal cell was mapped onto each one of the pile's 36 layers using appropriate trigonometric functions; each line segment obtained in any column of any layer of the pile after this mapping was compared with the

line segment (which it partly overlapped) retained from the previous "shift position" mapping, and the longer one was retained so that by the end of the whole tremor cycle only the best fit for every single stimulus line segment was still to be found in the pile. Once the whole tremor cycle was over, and once all stimulus line segments had been optimally mapped into their "templates", the vertical inhibition scheme described in detail in Chapter V was applied (using appropriate trigonometric functions again) causing the stimulus line segments to be identified with their respective positions, orientations and sizes.

The simulated pile was found to be highly efficient. We experimented with all sorts of line segments varying in position and/or orientation and/or size. We restricted our choice of input line segments to those whose orientations were multiples of five degrees (plus or minus one degree) to avoid ambiguous situations where a straight line segment, because its orientation is half way between two successive "orientation templates" in the pile, does not lead to a single longest line segment in one given layer but leads to two "longest" line segments overlapping in two successive layers of the pile (i.e. a curve specific outcome). This "orientational" resolution problem is very similar to the "positional" resolution problem which was solved by introducing (translatory) tremor, so in order to solve it we thought that we might introduce rotatory tremor. We however decided to give a rather low priority to this task.

The line segment detection scheme proved to be totally reliable for any line segment with a size greater than six receptive fields, but as shorter and shorter line segments were considered the scheme became, not very surprisingly, less and less reliable to the point of being hopeless for line segments extending over only two receptive fields. By "reliable" we mean that for any stimulus line segment the line detection scheme would come up with a single line segment in a particular layer of the pile. The problem with shorter line segments was that they yielded successions of overlapping "longest" line segments in different layers instead of yielding single line segments on the single layers representing their actual orientations. This is due to the fact that for any chosen orientational resolution a certain minimum length of line segment has to be reached in order to allow the given mapping criteria within the pile to make specific orientation detection possible. For instance with a resolution of 45 degrees a system can easily be specific about vertical and horizontal line segments of five units of length, but with a resolution of one degree a system can easily confuse a vertical (0 degree) line segment five units long and the same line segment tilted by one degree.

The point is that the finer the pile's resolution is, the longer line segments have to be in order to be uniquely specified. It seems that for an orientational resolution of five degrees, and given the mapping criteria of our pile, it is not safe to go for straight line segments which are shorter than six receptive fields. It is important to realise that this means only 24 retinal cells, a microscopic amount for the human eye.

Besides the few troubles with short line segments, experiments with the line segment detection pile were uneventful, all stimulus line segments yielding clear cut acurate results. Even curved line segments yielded the expected results, the pile being very sensitive to them. Figure C-4 shows the actual result obtained with the stimulus line segments whose mapping on the retina was shown in Figure C-1, or rather in Figure C-3 if one wants to consider tremor. Figure C-4 shows on the left the 36 layers of the pile with arrows pointing to those layers where line segments were left after the vertical inhibition was carried out, and these layers are shown on the right with their respective line segments; the top layer of the pile stands for 0 and the bottom one for 175 . Particularly interesting are the two partly overlapping line segments detected in two successive layers at the bottom of the pile and corresponding to the right hand side of the stimulus triangle; if one looks back at Figure C-1 this curve specific detection should appear far from surprising.

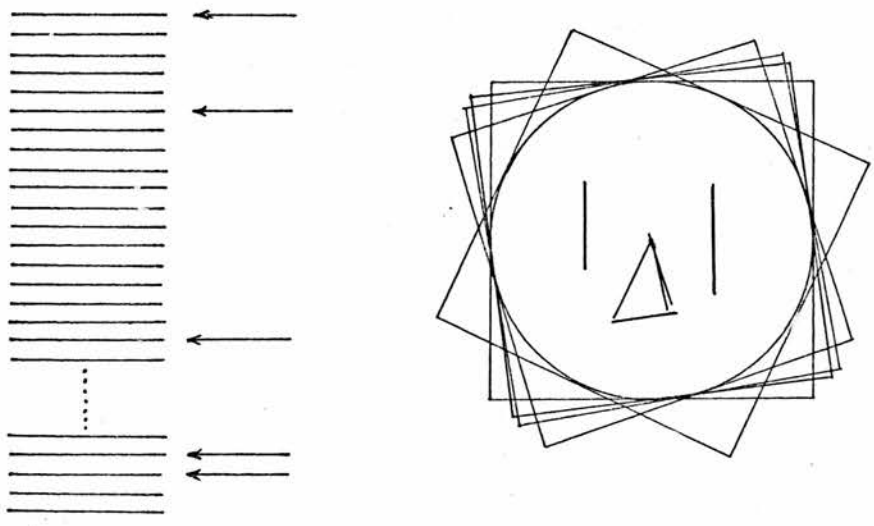


FIGURE C-4. Typical outcome of line segment detection in the twisted pile.

With the line detection pile successfully simulated we moved to simulating the computation of the transistence values presiding over object, sub-object, and background

differentiation. Very little need be said concerning the computerised version of this part of the system since computer programs can follow the nerve net operation very closely and since the computerised version of the general-purpose pile within which these operations were carried out has already been presented.

The main problem in choosing interesting stimulus line segments on which to experiment with transistence detection and grouping strategies was that on top of having to be contained within the boundaries of the system's rather small retina the chosen line segments had to be small enough to leave some room for motion to take place from moment to moment. A satisfactory starting disposition of stimulus line segments was found to be the one shown in Figure C-5a. The three frames shown respectively in Figures C-5a, C-5b, and C-5c were presented to the system in three successive moments and the following results were obtained. At each moment the three line segments presented were successfully identified with their respective positions, orientations, and sizes. At the first moment all three line segments' transistence value both in absolute and relative terms were found to be ON; this was of course the desired result. The first moment computation could obviously not lead to any differentiation between the line segments. At the second moment the first interesting results were obtained. In absolute terms the transistence value of both the vertical and the horizontal line segments (which were moved together in the stimulus structure from the first to the second moment) was found to be ON while that of the oblique line segment was found to be STILL. And in relative terms the transistence value of the vertical and the horizontal line segments was found to be STILL while that of the oblique line segment was found to be ON (the horizontal line segment having been chosen as reference). This is of course the desired result, a result which is in fact over-sufficient to differentiate two "groups" of line segments (the differentiation being possible either on absolute or on relative grounds). In the third moment all line segments were found to be ON in absolute terms and STILL in relative terms, this latter verdict indicating quite clearly a common movement of all three line segments. This is again the expected result. OFF transistence values were also obtained as expected in both moment-2 and moment-3, but since they were not essential to the points being made they were simply disregarded in the above account of the results obtained.

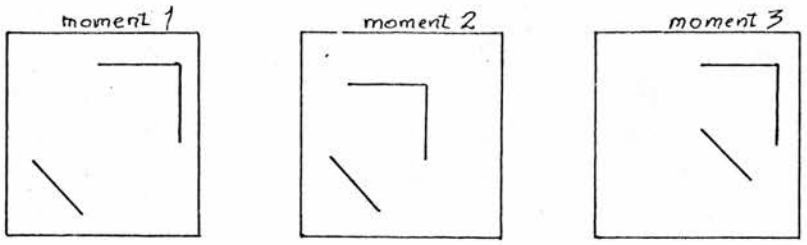


FIGURE C-5. First set of frames used to simulate groupings on running grounds.

A second series of sets of line segments, consisting of only two frames this time, was presented to the system to check on the efficiency of the relative transistence detection in cases where all line segments are moving in absolute terms but where some of the line segments move together in one way while the others move in another way. The two frames used are shown in Figure C-6. The first moment's analysis yielded the usual ON transistence value for all three line segments, and the second moment's analysis yielded the following results: in absolute terms all three line segments were found to be ON, which is hopeless for any differentiation, but in relative terms the horizontal and vertical line segments were found to be STILL while the oblique line segment was found to be ON, providing a perfect basis on which to separate the two "groups" moving in different ways.

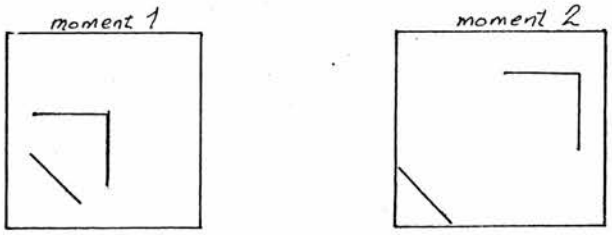


FIGURE C-6. Second set of frames used to simulate groupings on running grounds.

Finally the system was presented with a series of three frames allowing the full grouping procedure to be carried through, i.e. allowing a background, an object, and a sub-object to be defined on the basis of transistence values. The three frames are shown in Figure C-7. Being provided with these frames in three successive moments the system behaved as follows. In the first moment the usual ON values were detected. At the second moment the horizontal and vertical line segments were detected as being ON in absolute terms and STILL in relative terms while the oblique line segment was detected as being STILL in absolute terms and ON in relative terms. This allowed the system to distinguish between an object and a background. For the purpose of the present example we made the system send ON line segments in absolute terms to the object pile and STILL line segments in absolute terms to the background pile, but since the two object line segments were STILL in relative terms there was no evidence left on the basis of which a sub-object could be derived. With the third moment, however, such evidence became available: the oblique line segment remained STILL in absolute terms, and therefore had to be sent to the background pile again, while the vertical and horizontal line segments were also once again detected as being ON and sent to the object pile, but this time the vertical and horizontal line segments were not both STILL in relative terms, the horizontal one (being the reference) being detected as STILL and the vertical one being detected as ON. So on this basis the vertical line segment could be sent to the sub-object pile, completing the whole procedure.

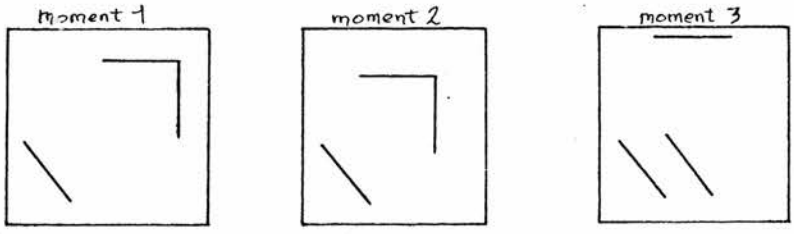


FIGURE C-7. Third set of frames used to simulate groupings on running grounds.

The simulation of this part of the system was therefore quite successful although it is rather regretable that we could not allow for more simultaneous line segments to be experimented with: ending up with single line segments in the background, object, and sub-object piles is a rather minimal achievement.

C.3 M-characterization and motion detection

The computer testing of all M-characterization and motion detection strategies was not completed; the parts of the system which have not been simulated yet are (1) the proportional size detection piles and their associated motion detection piles, (2) the whole of the Secondary System, (3) acceleration-deceleration detection piles. The reason why these parts of the system were given lower priority are that (1) acceleration-deceleration detection logically comes at the end of the whole motion detection process and is, with proportional size detection, just about the easiest type of motion to detect, (2) the Secondary System was successively used, as seen in Chapter VI, as a paradigm for designing experiments on human visual abilities and has therefore already been tested in some way, and (3) proportional size detection and its associated motions are, as mentioned in (1), easiest to achieve among all remaining detection tasks in the system. All remaining parts of the system's micro-structure have been simulated to some extent.

Assigning a global position (relative to the retina) to each one of the three sets of line segments found respectively in the background pile, the object pile, and the sub-object pile was the first process to be simulated. Implementing the (nerve net) global position detection strategy turned out to be straightforward enough but for the now familiar fact that since inter-layer "journeys" through a twisted pile are required by this strategy trigonometric functions introducing non-linearities have to be used to express these "journeys" in computer programs. However, this problem having already been solved elsewhere in the system, programs fulfilling all the requirements of the nerve net strategy were easily produced and tested. The usual retinal size (i.e. 16X16 positional receptive fields) was used for the simulation, so we suffered the usual restrictions on stimulus size. It is important to realise that such restrictions are more serious in the context of frozen feature detection (such as global position) than in the context of running feature detection. We nevertheless decided to accept the restrictions for a first experiment with our programs and we used, stimulus structures which are very similar to the ones which we used when simulating groupings on running

grounds (see last section). Not very surprisingly, problems of retinal resolution were encountered once again. These can be understood by realising that while on the one hand the strategy aims at finding the retinal point where signals from all line segments pass on their respective outward journeys, on the other hand these outward journeys are made along a very finite set of different directions (i.e. twice the number of layers in the twisted pile) and the progression in each one of these directions consists of discrete "jumps" (i.e. one complete cell at a time). This means that the retinal point where signals from all line segments pass becomes most of the time a retinal region extending over variable numbers of retinal cells depending on the stimulus structures. These problems appeared more especially acute as the retina was small, but they were finally handled by different approximation strategies and consistent global positions were obtained. Figure C-8 shows a typical stimulus structure, i.e. a sequence of five frames each of which consists of three straight line segments (A,B, and C). As these frames are successively presented to our simulated system the line segments are firstly classified as being background, or object, or sub-object line segments; the respective classifications for each frame are given under each one in Figure C-8. Then each group of line segments in each frame is given its global position; this global position is indicated in Figure C-8 within the frames themselves by a small circle labelled with the letter(s) corresponding to the line segment(s) whose global position is represented.

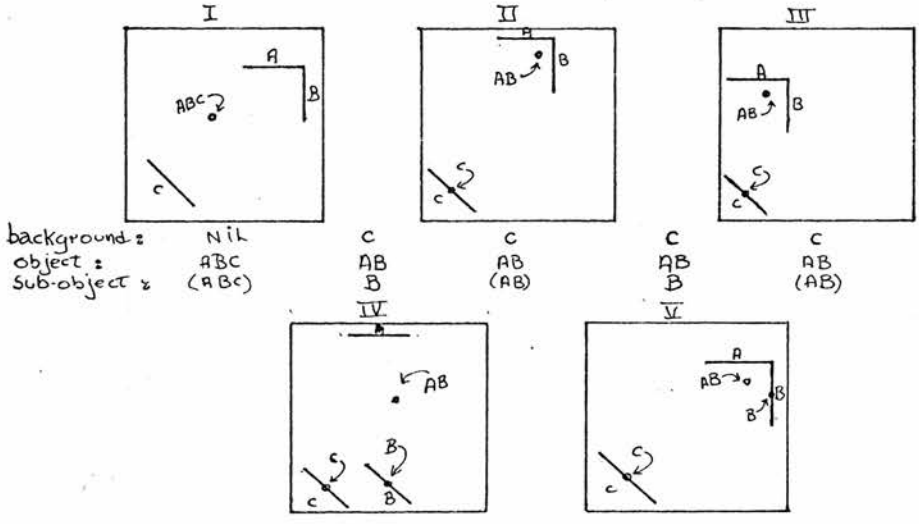


FIGURE C-8. Global positions obtained.

Having obtained global positions we turned to the simulation of global orientation detection. Once again the simulation was rather straightforward except for

expressing "journeys" through the twisted pile where most of the computing was to take place. The problem was solved in the usual way, i.e. trigonometric functions. The symmetry which the program was made to look for in order to assign global orientations was a rather weak one, but it was quite sufficient to provide us with adequate global orientations for the purpose of computing rotations. The program was first made to check on possible simple rotations of the set of line segments by matching the mapping of the current set of line segments in every layer of the pile with the mapping of the previous moment's set of line segments on the layer which was chosen to represent their global orientation. If no matching layer was found the orientation detection routine was triggered and it looked for the layer on which the line segments offer the most "economical" or "symmetrical" lay out. The criteria which were used in this search for "the best layer" are as follows.

1-Emphasis on axis of symmetry: on any given layer every two line segments which came from equally distant layers scored one point for this layer; if the given layer was the previous moment's winning layer two points instead of one were scored; and if the given layer was the top layer of the pile (i.e. the vertical layer) an extra one and a half points were scored.

2-Emphasis on orthogonality: on any given layer every line segment which came from eighteen layers away (i.e. 90 degrees) scored half a point for this layer; if the given layer was the previous moment's winning layer an extra half point was scored; and if the given layer was the "vertical" layer one extra point was scored.

3-Emphasis on retinal orientation: every line segment scored one point for its layer of origin; if this layer of origin was the previous moment's winning layer one extra point was scored; and if it was the "vertical" layer one extra point was also scored.

In the case of many layers having equal scores the closest layer to the previous moment's winning layer won. If there was no previous moment's winning layer the closest to the "vertical" layer won.

The size of the simulated retina restricted its use to rather trivial shapes, but symmetries, orthogonalities, and verticality were easily detected by the program and consistent global orientations were obtained.

Figure C-9 shows the detected global orientations for each one of the different sets of line segments shown with their respective global positions in Figure C-8.

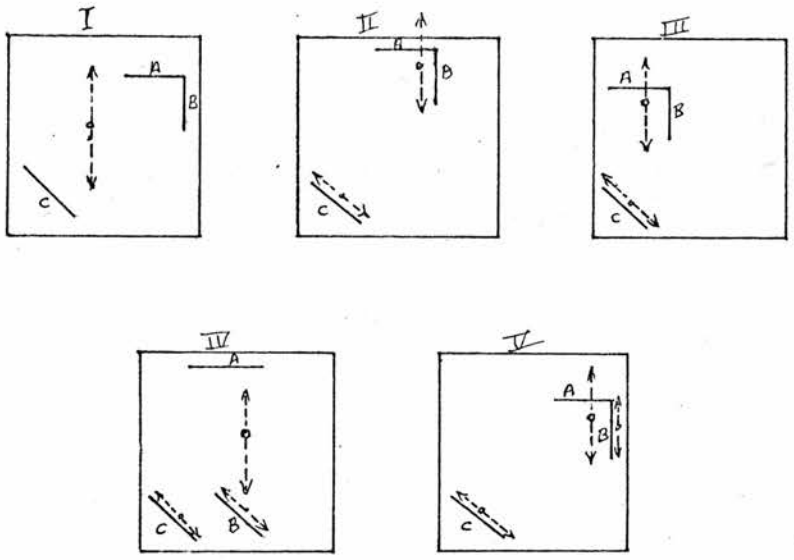


FIGURE C-9. Global orientations obtained.

Having obtained global positions and orientations of sets of line segments relative to the retina the next step was to simulate the detection of global positions and orientations of sets of line segments relative to other sets of line segments (i.e. object relative to background, and sub-object relative to object). The nerve net strategies for detecting both features were easily implemented, no major computer representation problem having to be tackled. We however had to restrict the computation of relative positions to taking into consideration the reference object's global position and orientation only, no program dealing with size detection having been written. The programs were tested on the five frame sequence portrayed in Figures C-8 and C-9, where the respective roles (i.e. background, object, sub-object) of the sets of line segments in each frame were interchanged in all possible interesting ways. The desired results were obtained in all cases.

Our last move in simulating the system's micro-structure was to have some actual motions detected. With the above described results we were ready to simulate the detection of six different motions out of the nine which our system is designed to compute on the basis of its frozen M-characterizing features; these six motions are the visual object's translation relative to the retina (1) and relative to the background (2) as well as its rotation relative to the retina (3) and relative to the background (4), and the sub-object's translation (5) and rotation (6) relative to the object. Computing these six types of motion actually involved minimal programming requirements.

Since TDU's had already been programmed we simply had to use similar programs in the context of our motion detection piles. As far as rotations are concerned the need for only two travelling OFF lines (clockwise and anti-clockwise lines) made the grouping of OFF and ON orientations into a rotational velocity almost trivial. Computing translations was a little more exacting but, as already mentioned, translatory movement is the running equivalent of line segment detection, and programs had already been written for line segment detection. Indeed the very fact that we already had a successful computer scheme for detecting line segments might have been regarded as sufficient evidence that our ideas about translatory movement detection were computationally sound. Nevertheless, the existing programs were reorganised and actual rotations and translations were obtained, instances of each one of the six different types mentioned above being successfully tackled.

Appendix D

Published paper

Perception, 1973, volume 2, pages 167-180

A new experimental paradigm for the investigation of the secondary system of human visual motion perception

C Lamontagne
Bionics Research Laboratory, School of Artificial Intelligence, University of Edinburgh
Received 12 July 1973

Abstract. An experimental paradigm is derived from a computational model of visual motion perception. The new family of phenomena that support this paradigm is presented in the context of the model. The basic phenomenon can be considered as being the apparent motion and the sustained eyetracking of a physically still object (relative to the subject) in the absence of any other object moving relative to the subject.

1 Introduction

In the course of developing a computational model of motion perception we came across a complete family of predictions which, when confirmed on human subjects, yielded an experimental paradigm for quite a detailed investigation of what has sometimes been referred to in the psychological literature (Gregory, 1966) as the eye-head system of motion perception. For reasons that will become quite apparent as the paper unfolds we will refer to this system as the *secondary system* of visual motion perception.

Since we basically wish to present the experimental paradigm mentioned above, we will try to restrict the discussion of our theoretical model to those features that are sufficient to let the reader grasp the general context within which this paradigm was elaborated.

2 Some features of the model

2.1 *Nature of the model*

Our model of motion perception is being developed along the following lines:
(i) We view the study of motion perception as an entry into visual perception as a whole. Although we are primarily concerned with building a model of motion perception, we are trying to design this model as an 'open' system, i.e. a system that can be used as a subsystem in a larger context at a later stage in our research.
(ii) We want to construct a system which performs at a human level of sophistication, excluding situations based on binocularity (i.e. our system is monocular) and those based on the retina's dual structure of a fovea and a periphery (i.e. our system has a homogeneous retina).
(iii) We want to build a *computational* model; that is we want to evaluate the processes which are candidates for some job in our system *in terms of their relevance to the computation that requires to be done.*

2.2 *The primary system of visual motion perception*

We started off by considering a visual system working on a 'discrete sampling' basis, and only equipped for the detection of different positions and different intensities of the 'beams' of light falling on a retina covering a very limited part of the environment. This retina can be described technically as a two-dimensional array of (differentially) light-sensitive units, each being specific to its particular position in the array.

By analysing positions and intensities (i.e. retinal data) a multitude of 'higher level' features can be derived. We can see two main types of features: the 'frozen' features which are derived exclusively from the set of retinal data gathered *during one sampling period,* and the 'running' features, which are derived from some *temporal succession* of some type of frozen feature. The purpose of our motion analysis system is clearly to compute the running features.

Our first step in designing the motion analysis system was to look for processes to compute very basic running features, so basic in fact that the actual type of frozen features involved would be irrelevant to the nature of the computation. In other words we wanted to deal first with those running features which are the highest in the hierarchy, that is those features which are to be found in any motion no matter what is moving.

For example, position is a frozen feature which can have many different values (e.g. position 1, position 2). If position 1 is an activated feature in one sampling period and is no longer activated in the next sampling period, we say that its 'running existence state' is 'off'. There are four different running existence states: 'on', 'off', 'still existing', and 'still not existing'. They represent the most basic type of running features. Now we are interested in motion, that is the nature of the change in value of some frozen feature. But the running existence state gives us only *local* information, that is it informs us of what is happening to a particular value of the frozen feature, and what we want is more global information, that is from what value to what value has 'existence' been transferred. Well, if existence has been transferred it means that some value has been 'turned off' and some other one has been 'turned on', so what we have to find is some scheme for combining these relevant local outcomes for computing velocity—the second most basic type of running feature.

So we started building our motion analysis system by designing the two processes which would compute our two basic running features. We called the first one a change detection unit (CDU), and the second one a velocity detection unit (VDU). We then had structures specific enough to allow us to start exchanging requirements with the frozen features analysis system. There were very few frozen features on which we could try our CDUs and VDUs at that time. We disregarded the potential running features of variations of intensity, and didn't attack then the fascinating prospect of jumping one level up and considering velocity as a frozen feature recursively pushed into the CDUs and VDUs to get at any desired derivative. We were then left with one possible load for our basic structures: positions of atomic units of some light intensity on the retina. Out of these, and following requirements brought from all sorts of sources, we started creating higher-level frozen features, matching each one of them with an adequate motion detection unit (MDU), that is a *particular* 'coupling' of some CDUs with a VDU, to allow the computing of their 'running' features when desired. When the requirements for new frozen features stopped pouring in, we had a visual system consisting of nine different types of frozen features and their nine corresponding MDUs. As expected, the structural generality of the CDU and of the VDU stood up to the pressure all the way, and three types of 'coupling' were found to be sufficient to design all the nine MDUs. We therefore had a very homogeneous structure. The nine types of frozen features are as follows: global position (1), global orientation (2), and global size (3) of some visual object *relative to the retina*; global position (4), global orientation (5), and global size (6) of some visual object *relative to some other visual object*; and *shape* of some visual object in terms of relative position (7), relative orientation (8), and relative size (9) of the local elements it consists of. 'Visual object' refers to some set of atomic features (retinal atomic positions and intensities) grouped under some criterion or other; hereafter we will refer to it simply as an 'object'. An important

point to be made about the nine MDUs is that they work in parallel, that is they compute the nine different types of motion at the same time. However, a heavy restraint on the system is that only one object can be analysed at a time—at least as far as the VDUs are concerned. This constraint, together with a few other ones of the same type elsewhere in the visual system, brought us to realise that if our system cannot 'see' everything that it ought to see 'at one glance', we have to provide it with some means of going from one point of interest to the other. We therefore designed two types of 'saccadic' systems to allow the visual system to extract from the visual world whatever sample it felt like analysing. One system performs 'physical' saccades by angular displacement of the eye (relative to the rest of the organism) so that the field of view (in part or in whole) can be changed, and the other system performs 'attentional' saccades by 'displacement' of computational power (within the same field of view or not) so that the focus of direct analysis can be changed. The two systems obviously work in close collaboration.

Now as the concept of 'physical saccade' goes along with the concept of 'physical retina' (the physical saccade involving displacement of the physical retina) we created the concept of 'attentional retina' to go along with the concept of 'attentional saccade'. Physical saccades have the nasty side-effect of creating a motion of the environment relative to the physical retina, so in order to avoid any useless computation by our motion analysis system, we decided that every saccadic command sent to the oculo-motor system would be accompanied by a general inhibition of the retinal input for the duration of the saccade, and that the saccade should be carried out as quickly as possible. Similarly, attentional saccades create disturbing side effects for motion perception by changing the positions of objects relative to the attentional retina. We eliminated these side effects by deciding that every attentional saccade would trigger off a re-set of all CDUs in our motion detection system.

At this point there was only one other type of situation (involving motion) to be discussed before closing the question of two-dimensional motion detection. Until then we had assumed that the object's motion relative to the retina was equivalent to its motion relative to the organism ('organism' referring to the physical system to which the eye is fixed), except maybe for those motions created by the activation of the saccadic systems mentioned above, in which case we were careful not to allow our system to compute motion. However, some major requirements forced us to provide our visual system with an *eyetracking system*, thereby creating a situation where this equivalence of frames of reference would be broken. One of the requirements was coming from the frozen features analysis system and was based on the impossibility of extracting some critical features of objects which are involved in some motion *relative to the retina*. One solution was to change the sampling period of the visual system according to the speed of the object under analysis, but a much easier one was to null the object's velocity relative to the retina by having the retina (i.e. the eye) 'track' the object. The problem with eyetracking was that by involving a motion of the retina relative to the organism it created a situation where our motion analysis system was totally incapable of computing 'positional' velocities relative to the organism (since it worked on a retinal basis). We therefore had to work out some system that would compute motion in this new type of situation. In fact we needed it for two reasons: first for 'seeing' at all times objects moving relative to the organism, and second to drive the eyetracking itself. But as the required system was taking shape, we realised that it had nothing whatsoever to do with the familiar MDU structure. In fact, it turned out that no motion as such had to be worked out; only some combining of motions that had already been 'worked out' by the old MDUs was required. So this new system was in fact forming a new layer of its own on top of the primitive MDUs, setting its roots deep into them ready to extract all the required information. In order to crystallise

these structural relations we called the new system the secondary system of visual motion perception and we put the nine MDUs into a big box which we called the primary system of visual motion perception. We will hereafter refer to these systems simply as the 'primary system' and the 'secondary system'.

2.3 *The secondary system of visual motion perception*

The general idea behind the secondary system is that in order to perform efficient eyetracking, and in order to allow an optimum awareness within the visual system while the eyetracking is under way, we need a system which computes velocities relative to the organism whenever some eyetracking is going on.

Our first version of the secondary system worked in the following way. When some object is to be tracked by the eye, this object's velocity relative to the retina (as worked out by the primary system) is sent to the secondary system. In the 'initialising' phase of the eyetracking the task of the secondary system is quite straightforward: since no tracking was taking place just before, the object's velocity relative to the retina is equivalent to its velocity relative to the organism, and the secondary system only has to put the label 'velocity relative to the organism' on its input to transform it into the desired output. Now this output is sent (i) to other parts of the visual system for further analysis; (ii) to the oculomotor system where the eyetracking is triggered off; and (iii) back to the secondary system itself as an input for the next moment. When the next moment comes, since the eyetracking is on the way, the object's velocity relative to the retina is no longer equivalent to its velocity relative to the organism, so that this time the secondary system has got a little more work to do to get its output computed. The way in which the object's velocity relative to the organism is computed in this case is that the secondary system carries out a vectorial analysis on its two input velocities, namely the object's velocity relative to the retina as provided by the primary system, and the eyetracking velocity as provided by the secondary system itself through its own latest output (third item in the list above). The output generated in this way by the secondary system is then sent to the same three destinations as in the 'initialising' phase, and the procedure that we have just finished describing is repeated until either the object can no longer be tracked or the visual system dismisses it as the focus of attention. It might be worth noting that the secondary system can be considered at all times as the vectorial analysis system that we just described, the 'initialising' phase being a case where one of the two input velocities (namely the eyetracking velocity) is a null velocity. The schema in figure 1 might advantageously express what we tried to describe verbally in the above paragraph.

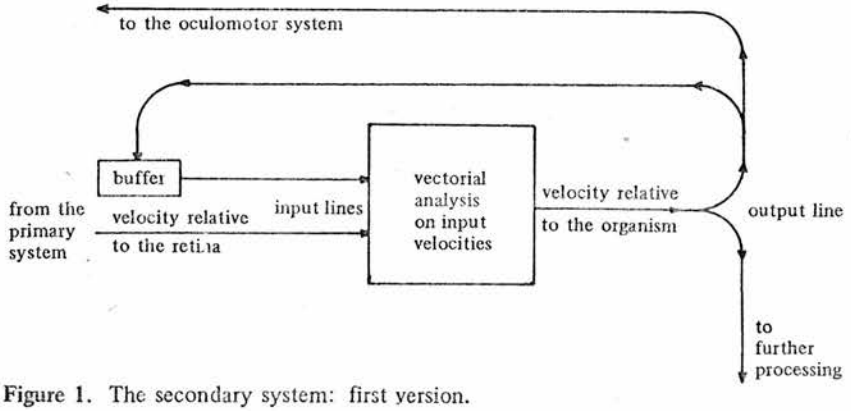


Figure 1. The secondary system: first version.

This secondary system allowed sustained eyetracking in cases of nonuniform motion as well as in cases of uniform motion. However, we noticed that in the case of tracking an object in uniform motion, since the tracked object is at all times kept 'still' relative to the retina, the primary system is only involved in computing immobility and the secondary system is reduced to carrying out vectorial analysis involving only one non-null velocity, namely the eyetracking velocity. This means that trivial computations are monopolising both the primary system and the secondary system (since they can only work on a single object's velocity at a time, however trivial the computations are) whenever the tracked object is in uniform motion. This is fair enough if one realises that in most cases one cannot tell when the motion will cease to be uniform and that, when it does, the full power of both systems will be needed to make the necessary corrections. Nevertheless, we came to the conclusion that we ought to provide the secondary system with an auxiliary system which could take care of the 'simple' computing required when the tracked object is in uniform motion; this would leave the visual system free to investigate other moving objects in the field of view while automatically tracking the object it was initially interested in. We therefore decided to split the secondary system into two distinct parts: the main secondary system (MSS), which is a structural replica of the former secondary system, and the auxiliary secondary system (ASS), which is the new structure for handling the eyetracking of objects in uniform motion. This new development was a complication, but the power of the secondary system was considerably increased.

Here is how the new secondary system works. The information provided to the secondary system by the primary system is divided into two groups: the *null velocities* ('on' and 'still' features computed by some CDU) which are sent to the ASS, and the *non-null velocities* (computed by some VDU) which are sent to the MSS. The MSS is used to 'initialise' the ASS in the following way. Whenever the visual system decides to track an object, this object's velocity (as computed by the primary system) is sent to the MSS (it has to be a non-null velocity or else no tracking would be required). The MSS then carries out its vectorial analysis and sends the result (i) to other parts of the visual system for further analysis; (ii) back to itself as an input for the next moment; and (iii) to the ASS. The ASS swallows the input and, without bringing any alteration to it, (i) sends it to the oculomotor system where the eyetracking is triggered off; and (ii) feeds it back to itself as input for the next moment. Apart from its own output the moment before (which we will call the 'local' input), the ASS can receive only one of two possible inputs (which we will call the 'foreign' inputs) at every moment: it is either the output of the MSS (when a new tracking velocity is required), or a null-velocity signal from the primary system (when the motion of the tracked object remains uniform). When the foreign input is the MSS's output, then the ASS ignores the local input (from its own latest output) and goes through the same routine as it did when 'initialised' [(i) and (ii) above]. However, when the foreign input is the null-velocity signal from the primary system, then the ASS takes the *local* input and (i) sends it up as a command to the oculomotor system; (ii) sends it to the other parts of the visual system for further processing; and (iii) feeds it back to itself as local input for the next moment. As long as the tracked object remains in uniform motion, the ASS can handle it very well on its own by going through the loop we just described. A representation of the main features of the new secondary system is given in figure 2.

An interesting point is that this new secondary system gives our system the ability to 'shift' the eyetracking from one object moving at a given velocity to a different object moving at a different velocity without having to break the eyetracking in the process. This is actually done in two steps: first the computational power of both the primary system's VDUs and the MSS are transferred from the currently tracked object to

some other moving object in the field of view, leaving to the ASS the task of handling the eyetracking itself. This transfer is achieved by what we called earlier an attentional saccade. Then the second step consists of transferring the eyetracking itself by 're-initialising' the ASS with the velocity of the new object (as worked out by the MSS after completion of the first step). This transfer will generally be accompanied by a physical saccade for reasons that will be made clearer in the paragraphs that follow. It is important to realise here that if the tracked object's velocity happens to change *after* the first step has been completed, and *before* the second one is completed, then the eyetracking breaks down (since the ASS alone can only cope with uniform motion). But it is also important to realise that this critical inter-step period can be made very short indeed.

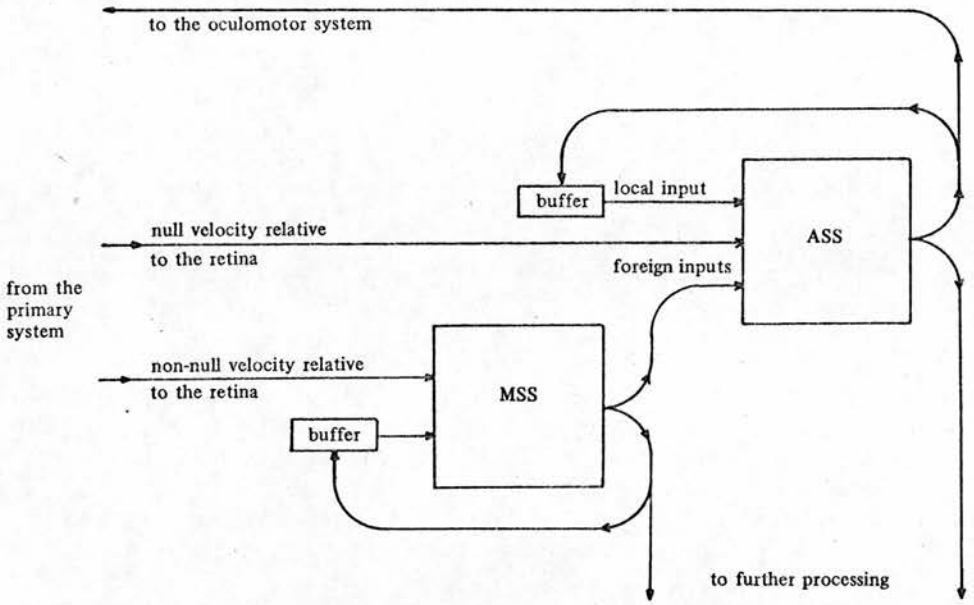


Figure 2. The secondary system: second version.

Now this business of changing the eye-tracking velocity, whether it appears in the context of tracking a single object in nonuniform motion or in the context of transferring the eyetracking from one object to another, raises an important problem which we have not yet dealt with. When we take a closer look at how our present system deals with changes of eyetracking velocity, we realise that these changes are inevitably brought one 'moment' after the object to be tracked has undergone the change in velocity (for the obvious reason that the change has to be detected before the system can start coping with it). The consequence of this delay is that the eye either loses or gains ground on the object (depending on the type of change involved) every time a change of velocity occurs, and, since the visual field of the eye is a limited one, a succession of such changes might progressively 'push' the object out of the visual field. We concluded from this that, whenever a change of eyetracking velocity is required to *match* the tracked object's velocity, an 'extra' motion is also required to *catch up* with the object. This 'extra' motion can in fact be carried out simply as a physical saccade when the new eyetracking velocity starts operating. So along these lines we decided to allow saccades within the eyetracking process itself: the commands for saccades would simply be combined with the commands for tracking whenever required. This new setup allowed the visual system to keep the tracked object more or less in the same spot on the retina throughout the whole tracking. To optimise

the situation we also decided to make it a rule that the tracked object should be kept in the most central region of the retina.

Now in this context we want to mention the fact that besides physical saccades we also had to allow attentional saccades (basically to keep the object in the centre of the attentional retina). This was an important move, because there was a great danger of interrupting the continuity of processing required to carry on the eyetracking since attentional saccades involve a reset of all the CDUs and since physical saccades involve an inhibition of the visual input. We lessened the danger considerably by deciding that the secondary system would consider 'on' signals computed by the primary system's CDUs as having the same velocity status (i.e. null velocity, or immobility) as the 'still' signals. This is the reason why 'on' signals were included under the label 'null velocity' when we described how the second version of the secondary system worked.

This concludes our discussion of the basic structure of the secondary system, but until now very little has been said concerning the actual type of motion that this system will have to deal with. What we do know is that the input to the secondary system comes from the primary system, so, since this system has a repertoire of nine different types of motions, the question is now how many of these are given access to the secondary system. If one considers that the basic requirement behind eyetracking, and consequently behind the secondary system, concerned only motions relative to the retina, then six types of motion are immediately dismissed by the fact that there are only three MDUs concerned with motion relative to the retina. We are therefore left with those motions based on changes of position, of orientation, and of size relative to the retina. Ideally, each of them should be accounted for by some appropriate tracking system (and consequently by the secondary system) but the problem involved in setting up tracking schemes for changes of size (which required a retina with adjustable size, or some zoom-lens system) and changes of orientation (which required a retina that could rotate clockwise and anticlockwise), without mentioning the problem involved by having many tracking systems in operation at the same time, encouraged us to consider the tracking of changing positions only. This meant that our secondary system would be built around a single MDU: the MDU concerned with changes of global position relative to the retina. Figure 3 provides a representation of the secondary system component of the full system sketched in section 2.2.

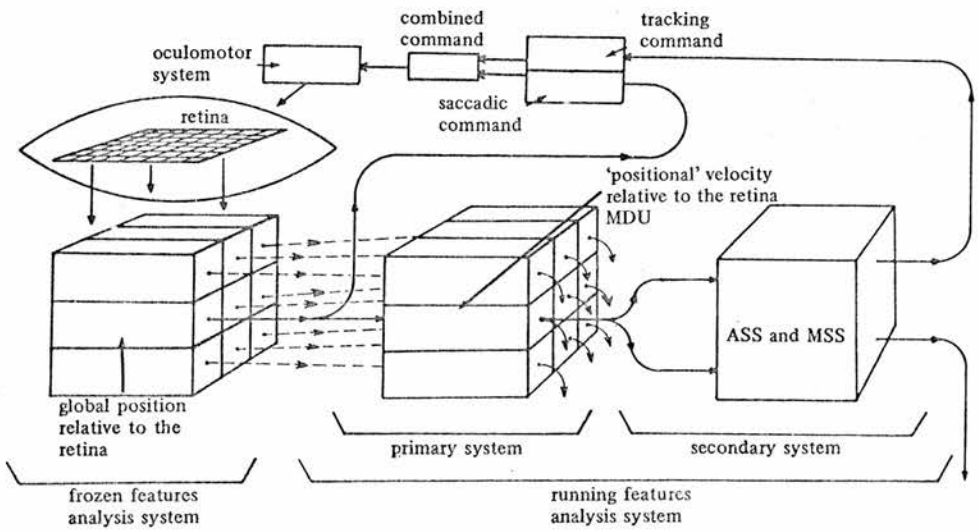


Figure 3. The secondary system and its context.

3 The predictions behind the paradigm

When we were designing the first version of the secondary system (see section 2.3) of our model of motion perception, it occurred to us that there existed a physical setup for accurately controlling the *critical* stimulus parameters underlying our model's analysis of motion. As the model developed, we found that this setup offered enough diversity to enable us to check on almost every ability of our system. Every time a new ability was given to the model, we tried to design a way in which we could use the experimental setup to check it. The aim of this exercise was obviously not to find an experimental scheme for studying our own model—what we had in mind was to use the experimental setup as a complete paradigm for the investigation of the human secondary system, to check the validity of our system as a model of it.

In the following sections each step of the design of this paradigm will be related to some part of the model. Also, each section will contain the results of those controls which have been carried out on human subjects to check the validity of that part of the model.

3.1 *The basic phenomenon*

In our model, once the eyetracking has been 'initialised', the secondary system only needs *null-velocity* inputs from the primary system in order to continue to *track* and *see* an object in uniform motion (see section 2.3). A situation that satisfies only these critical needs would be a situation where, once some tracking has been 'initialised', the tracked object is physically motionless relative to the organism as well as relative to the retina. If this situation could be found we would have a case of apparent motion where in fact a *stationary* object is being *eyetracked* as well as *seen in motion*!

The idea which we used to obtain such a setup is the following one. If our model's eye is made to track a target moving along under a stationary row of *identical* and *equidistant* objects, and if this row of objects is illuminated stroboscopically at a flash rate which is similar to the target's rate of 'passing under' the successive objects in the row, then we have a situation where the objects of the row are motionless for the primary system (i.e. relative to the retina). The reason for this is quite simply that, since every flash of light only occurs when the eyes have moved one inter-object distance along the row (i.e. the same set of retinal 'spots' always receives the row of objects) and since all objects in the row are identical, there is no retinal clue left to infer displacement.

In such a situation our system will 'see' the elements of the row moving along with its eye and the initial tracking target, and if it decides to transfer the eyetracking from this initial target to one of the elements of the row it will be able to carry on tracking this element and observe its motion until the very end of the row.

We therefore carried out an experiment where we used a row of 250 equidistant black dots as our row of objects, and a small spot of light running over the row of dots as our tracking target (see figure 4). We used 30 minutes of arc as interdot distance, 50 Hz as stroboscopic flash rate, and consequently 25 degrees per second as the speed of our tracking target.

We observed that as soon as our eyetracking was 'initialised' (using the moving spot of light) the whole row of black dots jumped into motion. We then concentrated on the black dot which the spot of light seemed to be riding and we took the spot of

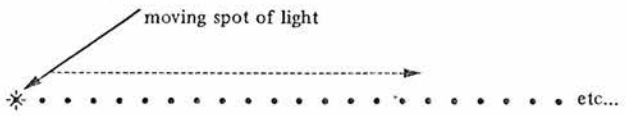


Figure 4. The basic stimulus setup.

light away: the tracking continued smoothly, and the black dot appeared to be moving as before! We were then well 'locked' in a tracking loop which was taking us happily and 'unconsciously' from one *stationary* dot to the next. With a little practice we found out that we could trigger 'the phenomenon' extremely easily, using the moving spot for a fraction of a second only, and track to the very end of the row. Although the phenomenon was extremely convincing by itself, we nevertheless recorded (by EOG) a few eye movements to make sure that eyetracking was actually taking place after the withdrawal of the initial tracking target. The results clearly showed that eyetracking was indeed taking place. Hereafter 'the phenomenon' will refer to the sustained tracking and the consequently perceived motion of the elements of the row *once the initial tracking stimulus has been removed.*

The phenomenon can be experienced under quite a wide range of frequencies of stroboscopic lighting. For example, in the case of our first experimental setup where the interelement distance was 30 minutes of arc, the phenomenon could be obtained under flash rates chosen anywhere between 10 Hz and 150 Hz.

Theoretically the *upper limit* of the range of 'permissible' flash rates (given an interelement distance) depends directly and only on the physical eyetracking mechanism's speed limit: if the flash rate (combined with the given interelement distance) requires a critical tracking speed which exceeds the power of the tracking system, then clearly the phenomenon cannot be obtained. To avoid a common misunderstanding it might be worth stressing the fact that the flicker fusion frequency of the visual system involved is not a critical factor in setting the upper limit of the range of 'permissible' flash rates. One might indeed fall into the natural trap of equating continuous 'perceptual' lighting and continuous 'physical' lighting; we therefore stress that the phenomenon is produced on the basis of a discontinuous *physical* lighting *only* and that consequently the elements of the row don't have to be *perceived* as discontinuously lit to allow it. There is therefore no surprise in realising that the phenomenon can be obtained easily with flash rates well *above* fusion.

On the other hand, the theoretical *lower limit* of the range of 'permissible' flash rates depends quite a lot on 'perceived' discontinuity. Here the matter is a little tricky, but the basic idea is the following one. As the flash rate is brought below fusion (i.e. below perceptual continuity) and reaches a point where the elements of the row don't trigger 'still' signals anymore, the phenomenon is still allowed for some lower frequencies since the secondary system accepts 'on' signals as meaning 'immobility' (see section 2.3); but as still lower flash rates are selected the 'on' signals are pushed further and further towards the edge of the sampling period until the stage is reached where the time interval between flashes is so long that a complete sampling period is deprived of its 'on' signal, thereby breaking the continuity of tracking by dropping below the lowest possible 'permissible' flash rate.

The fact that the phenomenon could be obtained at flash rates much *below* fusion therefore constitutes an experimental outcome in favour of the hypothesis that the secondary system works on 'on' signals as well as on 'still' signals. This hypothesis is also supported by results obtained by Gregory (1958) in an experiment in which subjects were asked to track a physically moving self-luminous target in a room illuminated stroboscopically at very low frequency (5 Hz). Gregory's subjects reported an apparent motion of the room along with the tracked target. However, in such a setup, the apparent motion is continuously interrupted by the 'off' phase of the lighting cycle, and although the objects in the room can periodically be seen as moving with the eyes, they nevertheless move away from the centre of the retina and are not replaced by exact copies of themselves in the retinal spot that they were occupying previously. For these two reasons the setup allows neither the eyetracking

nor the apparent motion *in the absence of the self-luminous tracking target.* Although, in our context, this experiment provides evidence which is more partial than the evidence offered by the phenomenon on the treatment of 'on' signals, it is nevertheless important.

3.2 *More critical eyetracking speeds to obtain the phenomenon*

We saw in section 3.1 that a critical eyetracking speed (in degrees per second) for eliciting the phenomenon can be derived from multiplying interelement distance (in degrees) by flash rate (in Hz). We will now consider the critical speed worked out this way as the *basic* critical eyetracking speed, and we will discuss two ways of deriving from it *new* critical speeds (for a given interelement distance and a given flash rate).

First, any multiple of the basic critical speed is itself a critical speed (its actual 'permissibility' depending on whether or not it exceeds the power of the eyetracking mechanism; see section 3.1 on this issue). The reason for this can be grasped by realising that any speed of eyetracking which preserves the immobility of the elements of the row (relative to the retina) is a critical frequency, and that given a flash rate and an interelement distance any multiple of the basic critical tracking speed preserves this 'immobility'. Obviously the perceived speed of the apparent motion is expected to conform with each different critical tracking speed.

Secondly, submultiples of the basic critical speed are themselves critical speeds. The basic idea behind this theoretical expectation is here again that the submultiples of the critical speed preserve the immobility of the elements of the row relative to the retina although they create singular side effects. Let us take for instance a situation where we have a row of black dots 30 minutes of arc apart on a white sheet of paper under a flash rate of 100 Hz, and where the initial tracking target is set at a speed of 25 degrees per second (i.e. *half* of the basic critical eyetracking speed). As soon as the system 'initialises' its eyetracking using the moving target, the following stimulus pattern falls on the retina. A first flash of light projects the row of dots onto the retina in those spots marked 'F' (for first) in figure 5. Since the eye has travelled only half of the interdot distance when the second flash of light comes, the row of dots is projected onto the retina in those spots marked 'S' (for second). Now, when the third flash of light comes, since the eye has travelled a complete interdot distance since the first flash (i.e. twice half an interval), the row of dots is projected again onto the retina in those spots marked 'F', and when the fourth flash of light comes the row of dots is projected onto the retina in those spots marked 'S', and so on. So what happens is that *the same two sets of spots* on the retina are successively and repeatedly exposed to the black dots. This creates a situation where, if the flash rate is such that two successive flashes happen within one sampling period of the frozen features analysis system, our model's eye will experience

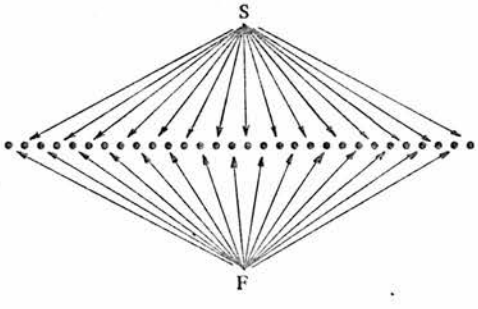


Figure 5. The 'doubling' effect at half the basic critical eyetracking speed.

the phenomenon on one single (simultaneous) row of dots where the dots are twice as close and twice as numerous as they would be in the case of eyetracking the same physical setup *at the basic critical speed.* Also, since in this situation 'every other flash' projects 'whiteness' on the black dots' retinal spots, our system will find the black dots 'sort of greyish'.

The general picture that emerges from this discussion is that when we divide the basic critical speed of eyetracking by some factor N (N being an integer) we get a critical eyetracking speed at which the model perceives a row of N times as many elements (with an interelement distance N times as short) as when it is tracking the same physical setup at the basic critical speed.

In the present state of our theory the phenomenon should be obtained for any value of N as long as the 'range limitations' discussed in section 3.1 are respected, and as long as the elements of the row receive enough light to allow a decent contrast. However, we feel somehow that the maximum value of N should otherwise be set at the point where the elements of the row get so close to each other that they form a perceptually continuous straight line (we want to allow the phenomenon for partially overlapping objects).

Now the following results were obtained from preliminary controls made on human subjects:

(i) Multiples of basic critical eyetracking speed were observed to be critical speeds themselves.

(ii) Submultiples were also observed to be critical speeds; the change in the number of elements of the row and in the interelement distance as a function of N was obvious. The highest 'N' which we managed to reach with the particular rows we were using was $N = 4$.

A rather obvious point is that the discussion of how *critical tracking speeds* can be worked out for a given flash rate and a given interelement distance is also valid (at least as far as the general idea of multiples and submultiples is concerned) for deriving how *critical flashing rates* can be worked out for a given tracking speed and a given interelement distance, as well as for deriving how *critical interelement distances* can be worked out for a given tracking speed and a given flash rate. This is hardly surprising since these three parameters are so much related in the production of the phenomenon; but we felt that the fact was worth noting all the same.

3.3 *Changes of velocity within the phenomenon*

Until now the phenomenon has allowed us to check on the ability to perform eyetracking in cases of uniform motion only. Let us now have a look at how the phenomenon can help us investigate the ability to perform eyetracking when nonuniform motion is involved.

Since nonuniform motion involves changes of velocity and since flash rate is one factor controlling tracking velocity in the context of the phenomenon, a relevant question would be: how would our model react to a progressive change of flash rate once the phenomenon has been triggered on some row of equidistant elements? Such a change would in fact create a situation where the flashes of light would occur *before* (if the flash rate is increased) or *after* (if the flash rate is decreased) the eye has travelled through a complete interelement distance. In this case, as opposed to the case of multiples or submultiples of the basic critical tracking speed (or flash rate), *different* spots on the retina are exposed to the elements of the row from moment to moment and therefore the elements' immobility is lost. This causes the primary system to detect changes of position relative to the retina and to compute a velocity from them. This velocity is then sent to the secondary system (in fact to

the MSS) where it is combined vectorially with the current eyetracking velocity in order to work out the new velocity of the element relative to the organism, velocity which becomes the next eyetracking speed. This means that in the event of a progressive change (obviously within certain quantitative limits) of flash rate our model automatically accelerates or decelerates accordingly (depending on the type of change), and the phenomenon is kept going.

For similar reasons, given a constant flash rate, progressive changes of interelement distances will trigger an automatic acceleration or deceleration of the eyetracking speed, thereby keeping the phenomenon going (see figure 6a).

Following the same type of argument we also reached the conclusion that the phenomenon will be kept going automatically in the event of a change of *direction* of eyetracking. In other words if the row of elements goes on straight for a while but then slants away at an angle of, for example, 20 degrees (see figure 6b), the model's eye should be able to take the tracked element over the bend without causing the phenomenon to break.

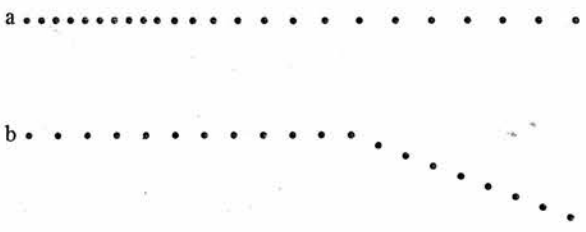


Figure 6. (a) Speed change stimulus; (b) direction change stimulus.

In those cases where changes of velocity are involved it becomes important to keep the attention of the visual system well focused on the tracked element, since, as we explained in section 2.3, the whole power of the system is required when changes in the tracked object's velocity occur.

The following results were obtained from preliminary controls made on human subjects:

(i) Given a row of equidistant black dots, the phenomenon was kept going under many rates of progressive increase and decrease of the flash rate. Corresponding perceptual impressions of acceleration and deceleration were observed.

(ii) Given a constant flash rate, the phenomenon was kept going as the tracked dot was taken over increasing (tracking left–right) and decreasing (tracking right–left) interdot distance (see figure 6a). Corresponding perceptual impressions of acceleration and deceleration were observed.

(iii) Given a 'bent' row of equidistant black dots under a constant flash rate, the phenomenon was kept going as the tracked dot was taken over the bend (see figure 6b).

(iv) The phenomenon broke down (i.e. subjects suddenly found themselves facing a set of stationary dots) in all three situations when the subject's attention was not focused on the tracked element when the changes came. This piece of data comes from reported 'impressions' only, since attention is a very hard thing to measure!

The fact that we can deal with changes of velocity has an interesting consequence as far as critical eyetracking speeds are concerned. It in fact means that there is no need to set the motion of the initial tracking target on the exact theoretical critical value: any approximation of this critical value which falls within the correction range of the secondary system will trigger the phenomenon (the corrections being automatically and immediately done by the secondary system). This makes the phenomenon much easier to obtain: for instance one only has to track one's own

finger while moving it at different speeds below the row of elements and a sufficient approximation of some critical speed (the basic one or some multiple) will soon be found.

3.4 *The effect of other types of motion on the phenomenon*
We have just seen in section 3.3 that changes of position of the tracked element relative to the retina affect the human system very much like they affect our model. In our model, however, these changes are the only ones to be coped with by the secondary system; we saw in section 2.3 that the only changes (detected by the primary system) which the secondary system considers are *changes in global position of the visual object* relative to the retina. So this means that no other type of velocity undergone by the tracked element will interfere with the phenomenon itself in the case of our model. Since in our model there are eight types of velocity other than positional velocity relative to the retina, we would have a long way to go to discuss the problem thoroughly. We therefore decided to illustrate the general technique of discussion in this context by chosing two out of these eight remaining velocities: 'rotational' velocity of the object relative to the retina (based on changes of orientation relative to the retina), and 'positional' velocity of the object *relative to some other object on the retina.*

To deal with rotational velocity relative to the retina we will consider a row of equidistant elements consisting of a set of line segments spread out in a *spatial* succession reproducing the *temporal* succession of those orientations involved in a rotational movement (see figure 7a). Since the object to be tracked is a line segment, the visual system has to work out a (global) position for it: in the present case this position has to be the centre of the line—if the line is to be tracked. Once the phenomenon is triggered in such a situation, our model 'sees' the tracked object as a *rotating* line involved in a *translatory* motion. The experiment was carried out with human subjects and the hypothesis was confirmed.

Now to deal with positional velocity of the tracked object *relative to another object* on the retina, we will consider a row of equidistant black dots, where each dot is 'framed' by a black circle which changes slightly its position relative to the dot from one dot to the next (see figure 7b). The interesting outcome here is that, once the phenomenon has been triggered, our model 'can see' the motion of the tracked dot relative to the circle *even when the direction of eyetracking is the opposite of the direction of the 'relative' motion,* and it can carry on with the tracking all the same. Again this was confirmed in an experiment with human subjects.

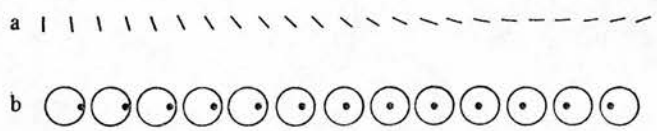


Figure 7. (a) Rotational motion stimulus; (b) relative translational motion stimulus.

3.5 *The transfer of eyetracking from one object to another*
We saw in section 2.3 how our model is able to transfer its attention (i.e. some of its computational power) to other objects when the tracked object is in uniform motion. We also saw how this allowed the transfer of the actual eyetracking from the 'old' object to the 'new one'. This capacity was in fact required for actually allowing the phenomenon in all the situations we have dealt with so far: we indeed had to go *from* the initial tracking target *to* some element of the row in every case. However, the process is under much better control in the following situation.

Let us consider a 'pile' of rows of dots where within each row the dots are equidistant, but where each row has got an interdot distance slightly shorter than the interdot distance of the row just below it (see figure 8).

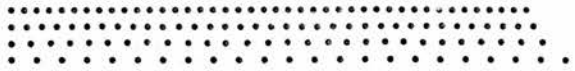


Figure 8. Eyetracking transfer stimulus.

If the phenomenon is triggered on the top row, our model's eye can then 'jump' from row to row right down to the bottom row of the pile, and climb back up again without ever causing the phenomenon to break. As we explained briefly in section 2.3, this involves, within the system, quite a complex interplay of attentional saccades and physical saccades, so this is quite a subtle exercise. However subtle the exercise, the human capacity to do the same was successfully controlled. After a little practice one could control the phenomenon beautifully, leaving the tracked element to the good care of the 'autopilot' (the ASS in our model) and observing other elements (moving at different velocities in the rows nearby and at the same velocity in the local row), or transferring the tracking itself to one of those observed elements (changing row if the observed element is in another row, remaining in the same row if the observed element is in it). An interesting outcome of this type of ability is that given a single row of equidistant dots, one can jump back and forth in this row allowing the phenomenon to be kept up indefinitely. We made a few eye-movement recordings of such a case, and the recordings clearly showed that after intervening saccades are made the eyetracking is automatically carried on as before.

3.6 *The experimental paradigm: conclusion*

We have seen in sections 3.1 to 3.5 how the occurrence and stability of the phenomenon could be used as a dependent variable in a reasonably rich context, that is a context where critical independent variables are potentially numerous and easy to control. By 'numerous' we obviously mean diversified enough to allow precise control of the whole repertoire of abilities embodied in the type of behaviour under investigation. However, if we have shown *qualitatively* how the critical parameters can be varied to check such and such an ability, we have said virtually nothing concerning the *quantitative* aspect of the problem. The reason for this silence is that we have not carried out any systematic experiment aimed explicitly at the quantitative investigation of some parameter or other. The experiments that we have carried out were only meant to check the relevance of some parameter in the context of some ability. The power of the paradigm is therefore far from being exhausted by the partial results reported in this paper, and we hope that, apart from having served its purpose in the context of our theoretical framework, it can stimulate other efforts in the field.

Acknowledgement. This work was done while the author was funded through a grant from the National Research Council of Canada. The author wishes to express his gratitude to his supervisor, Dr. J.A.M.Howe, for his help and encouragement, and also to those people of the Bionics Research Laboratory who have helped him complete this work.

References

Gregory, R. L., 1958, "Eye movement and the stability of the visual world", *Nature,* 182, 1214-1216.

Gregory, R. L., 1966, *Eye and Brain* (World University Library, London).

REFERENCES

ALLPORT, D.A. (1968) Phenomenal simultaneity and the perceptual moment hypothesis. Br. J. Psychol., 59, 4.

ATTNEAVE, F. (1968) Triangles as ambiguous figures. The American Journal of Psychology, 81, 3.

ATTNEAVE, F. (1971) Multistability in perception. Scientific American, December issue.

CLOWES, M.B. (1971) On seeing things. A.I. Journal, Spring issue.

DUNCKER, K. (1929) Translated in Ellis (1938), pp. 161-172.

ELLIS, W.D. (1938) A source book of Gestalt psychology. Kegan Paul, Trench, Trubner and Co.,Ltd., London.

FORGUS, R.H. (1966) Perception. McGraw-Hill, Toronto.

GAMOW, G. (1966) Thirty years that shook physics. Anchor Books Doubleday and Co., Inc., Garden City, New York.

GIBSON, J.J. (1957) Optical motions and transformations as stimuli for visual perception. Psychol. Rev., 64, 288-295.

GIBSON, J.J. (1966) The senses considered as perceptual systems. Houghton-Mifflin Co., Boston.

GREGORY, R.L. (1958) Eye movement and the stability of the visual world. Nature, 182.

GREGORY, R.L. (1966) Eye and brain. World University Library, London.

GREGORY, R.L. (1972) A look at biological and machine perception. In Machine Intelligence 7, Meltzer, B. and Michie, D. (Eds), Edinburgh University Press.

GRUSSER, O.-J. and GRUSSER-CORNEHLS, U. (1973) Neuronal mechanisms of visual movement perception and some psychophysical and behavioral correlations. Handbook of Sensory Physiology, volume VII/3A, Central Processing of Visual Information Part A, pp.333-429.

GUZMAN, A. (1968) Computer recognition of three-dimensional objects in a visual scene. Technical Report AI-TR-228, M.I.T. Artificial Intelligence Laboratory, Cambridge, Mass..

HARTLINE, H.K. (1938) The response of single optic nerve fibers in the vertebrate eye to illumination of the retina. American Journal of Physiology, 121, 400-415.

HARTLINE, H.K. (1940a) The nerve messages in the fibers of the visual pathway. Journal of the Optical Society of America, 30, 239-247.

HARTLINE, H.K. (1940b) The receptive fields of optic nerve fibers. American Journal of Physiology, 130, 690-699.

HARTMAN, L. (1923) Translated in Ellis (1938), pp.182-191.

HEYWOOD, S. (1973) Pursuing stationary dots: smooth eye movements and apparent movement. Perception, 2.

HUBEL, D.H. and WIESEL, T.N. (1962) Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. Journal of Physiology, 160, 106-154.

HUFFMAN, D.A. (1970) Impossible objects as nonsense sentences. In Machine Intelligence 6, Meltzer, B. and Michie, D. (Eds), Edinburgh University Press.

JOHANSSON, G. (1950) Configurations in event perception. Uppsala: Almkvist and Wiksell.

JOHANSSON, G. (1971a) Visual motion perception. Report 98, Department of Psychology, University of Uppsala.

JOHANSSON, G. (1971b) Visual perception of biological motion and a model for its analysis. Report 100, Department of Psychology, University of Uppsala.

JULESZ, B. (1971) Foundations of cyclopean perception. The University of Chicago Press.

KAPLAN, W. (1969) Kinetic disruption of optical texture: the perception of depth at an edge. Perception and Psychophysics, 6, 193-198.

KOHLER, W. (1920) Translated in Ellis (1938), pp.17-54.

KOHLER, W. (1947) Gestalt Psychology. Mentor Book, The New English Library Ltd., London.

KOLERS, P.A. (1972) Aspects of motion perception. Pergamon Press, Oxford.

KORN, A. (1974) Untersuchung Von Eigenschaften des Augenfolgesystems mit Hilfe Von Scheinbewegungen. Zeitschrift fur experimentelle und angewante Psychologie, Band XXI, Heft 3.

LAMONTAGNE, C. (1973) A new experimental paradigm for the investigation of the secondary system of human visual motion perception. Perception, 2, pp.167-180.

LEE, D.N. (1971) Binocular stereopsis without spatial disparity. Perception and Psychophysics, 9, 216-218.

LEE, D.N. (1972) Stimulus pairing in sequential phi motion. Perception, 1, 85-91.

LEE, D.N. (1974) Visual information during locomotion. In Perception: essays in honor of James J. Gibson, R.B. MacLeod and H.L. Pick, Jr. (Eds), Cornell University Press, Ithaca and London.

LETTVIN, J.Y. MATURANA, R.R. McCULLOCH, W.S. and PITTS, W.H. (1959) What the frog's eye tells the frog's brain. Proc. IRE, 47.

RETI, L. (1974) (Ed) The unknown Leonardo, McGraw-Hill Book Co. (UK) Ltd., London.

ROBERTS, L. (1963) Machine perception of three-dimensional solids. Technical Report 315, M.I.T. Lincoln Laboratory, Cambridge, Mass..

SCHOUTEN, J.F. (1967) Subjective stroboscopy and a model of visual movement detectors. In Models for the perception of speech and visual form, M.I.T. Press, Cambridge, Mass..

STOPER, A.E. (1967) Vision during pursuit movement: the role of oculomotor information. (Doctoral dissertation, Brandeis University.) Ann Arbor, Michigan: University microfilms No. 67-76, 579.

STOPER, A.E. (1973) Apparent motion of stimuli presented stroboscopically during pursuit movement of the eye. Perception and Psychophysics, 13, 2.

TERNUS, J. (1926) Translated in Ellis (1938), pp.149–160.

WALTZ, D.L. (1972) Generating semantic descriptions from drawings of scenes with shadows. AI TR–271, Artificial Intelligence Laboratory, M.I.T., Cambridge, Mass..

WERTHEIMER, M. (1924) Translated in Ellis (1938), pp.1–11.

YARBUS, A.L. (1967) Eye movements and vision. Plenum Press, New York.

GLOSSARY

ATOMIC VISUAL ENTITY (a.v.e.): Most primitive visual entity (v.e.), specifying the occurrence of light in a particular position at a particular moment on the physical retina. See p.18 for more details.

ATTENTIONAL RETINA: Conceptual retina loaded with all a.v.e.'s from that part of the physical retina which the system wishes to analyse in terms of a visual object, a sub-object, and a background. See p.108 for more details.

CHARACTERIZATION: Process by which a visual entity (v.e.) is given characteristics, or features, or attributes as a whole. See Section I.1 (p.16) for more details.

FROZEN FEATURE: Characteristic of a visual entity (v.e. obtained by grouping contemporary lower level visual entities. See pp. 21-22 for more details.

FROZEN GROUPING: Process of grouping together contemporary visual entities

M-CHARACTERIZATION: Process by which visual entities are characterized in terms of current values of multi-valued features, allowing transformations or movements to be detected. See p.32 for more details.

MULTI-VALUED FEATURE: Attribute or feature (of a v.e.) which takes at any moment any one of a variety of values in some domain (e.g. position, colour). See pp. 20-21 for more details.

PHYSICAL RETINA: Input device, consisting of an array of receptor-cells responding in an all-or-none manner to light intensities falling in particular positions at particular moments, yielding the system's most primitive visual entities, or atomic visual entities (a.v.e.'s). See Section I.1 (p.16) for more details.

PILES: Multi-dimensional arrays for computing and storing values of multi-valued features. See pp. 139-142 and 157 for more details. For twisted piles, see pp. 168-172, 178-191.

PRIMARY SYSTEM of visual motion detection: Part of the motion detection system which takes the retina (or the eye) as ultimate frame of reference.

RUNNING FEATURE: Characteristic of a visual entity obtained by grouping non-contemporary lower level visual entities. See pp. 22-25 for more details.

RUNNING GROUPING: Process of grouping together non-contemporary visual entities.

S-CHARACTERIZATION: Process by which visual entities are characterized in terms of single-valued features, specifying the visual entities' respective identities. See pp. 32-33 for more details.

SECONDARY SYSTEM of visual motion detection : Part of the motion detection system which takes the organism to which the eye (or retina) belongs as ultimate frame of reference. See pp. 238-251 for more details.

SINGLE-VALUED FEATURE: Attribute or feature (of a v.e.) which is either true or false in any given situation. Predicate. See p.20 for more details.

TRANSISTENCE: Multi-valued feature specifying the mode of existence of any value of any feature through two successive moments (its four possible values being ON, OFF, STILL THERE, and STILL NOT THERE). See pp. 133-134 for more details.

TRANSISTENCE DETECTION UNIT (TDU): Effective decision procedure (expressed in terms of a nerve net) to compute transistence. See pp. 145-149 for more details.

VELOCITY DETECTION UNIT (VDU): Effective decision procedure (expressed in terms of a nerve net) to compute velocity. See pp. 150-156 for more details.

VISUAL ENTITY (v.e.): Any entity which is given a characteristic as a whole at any point in the processing.

VISUAL OBJECT: Ultimate visual entity. Goal of the processing prior to actual motion detection.